# DEVELOPMENT AND VALIDATION OF THE SITUATIONAL TRUST SCALE FOR AUTOMATED DRIVING (STS-AD)

A Dissertation
Presented to
The Academic Faculty

by

Brittany E. Holthausen

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in Engineering Psychology in the
School of Psychology

Georgia Institute of Technology
August 2020

# DEVELOPMENT AND VALIDATION OF THE SITUATIONAL TRUST SCALE FOR AUTOMATED DRIVING (STS-AD)

Approved by:


Dr. Bruce N. Walker, Advisor
Schools of Psychology and Computing
*Georgia Institute of Technology*

Dr. Karen Feigh
School of Aerospace Engineering
*Georgia Institute of Technology*

Dr. Richard Catrambone
School of Psychology
*Georgia Institute of Technology*

Dr. M. David Keller
Human Systems Research and Development
*Naval Surface Warfare Center Dahlgren*

Dr. Jamie Gorman
School of Psychology
*Georgia Institute of Technology*



Date Approved: May 11, 2020

# ACKNOWLEDGEMENTS

I would like to thank my husband, who has continually supported the late nights and “please don’t talk to me” days of dissertating. I would like to thank my family for continuously encouraging me to pursue academic endeavors. I’d like to thank my pets, for continuously bringing me back to the present moment.

I would like to thank my advisor, Dr. Bruce Walker, for providing the guidance and encouragement I needed throughout this process. I would like to thank my committee for their insights, feedback, and encouragement.

I’d like to thank my lab mates who have always served as sounding boards for ideas and the best reviewers there are. To Rachel, for keeping our adjacent desks clean, our paperclip and sticky note wall, and for sharing her cooking and passion for life with all of us. To Keenan, for always taking time to go on winding research discussions. To Brianna, for your constant encouragement and for always inspiring me to eat more vegetables. To Zoe for including the STS-AD in her master’s thesis data collection so that I could have some high fidelity driving simulator data despite the stoppage of human-subjects research due to COVID-19. To Nadia, for keeping me on my toes with your fascinating research ideas. To Mengyao Li, although I was her mentor during her time at Georgia Tech, she taught me more than I taught her. I am so proud to have been a small part of your undergraduate career and continue to cheer you on in your graduate school endeavors. To all the other members of the Sonification lab, past and present for their unwavering support, wisdom, and friendship.

Lastly, I’d like to thank my collaborator in all of this, Philipp Wintersberger. He and I have worked together throughout the duration of our graduate degree pursuits.

We’ve put on three workshops together educating the AutoUI community on trust in automation and how to measure it, which culminated in the development of this work. Thanks, Philipp for providing insight into the computer science side of our work, listening to my psychology rants, and all the laughs along the way.

# TABLE OF CONTENTS


**ACKNOWLEDGEMENTS** **iii**
**LIST OF TABLES** **viii**
**LIST OF FIGURES** **ix**
**SUMMARY** **x**
**CHAPTER 1. Introduction and Motivation** **1**
**1.1 Motivation** **1**
**1.2 Thesis Statement** **1**
**1.3 Research Questions** **2**
**1.4 Summary of Studies** **2**
**CHAPTER 2. Background** **4**
**2.1 Automation** **4**
**2.2 Automated Vehicles** **5**
**2.3 Trust in Automation** **6**
2.3.1 Definition 7
**2.4 Models of Trust in Automation** **8**
2.4.1 Lee and See (2004) 8
2.4.2 Hoff and Bashir (2015) 10
**2.5 Trust in Automation Measurements** **14**
2.5.1 Self-Report Measures of Trust in Automation 15
2.5.2 Behavioral Measures of Trust in Automation 19
**CHAPTER 3. Development and Initial Online Validation of the Situational Trust Scale - Automated Driving (STS-AD)** **21**
**3.1 Background and Motivation** **21**
**3.2 Methodology** **22**
3.2.1 Participants 22
3.2.2 Materials 23
3.2.3 Procedure 25
**3.3 Results** **26**
3.3.1 Overview 27
3.3.2 Comparing STS-AD scores across videos 27
3.3.3 Construct Validity Evaluation 32
3.3.4 Inter-item Reliability Analysis 33
3.3.5 Situational Trust Scale Factor Analysis 34
**3.4 Discussion** **34**
**3.5 Conclusion** **36**
**CHAPTER 4. Methods for Driving Simulator Validation of the STS-AD** **37**
**4.1 Participants** **37**
**4.2 Apparatus** **39**
4.2.1 Driving Simulator 39
4.2.2 Non-Driving Related Task (NDRT) 39
**4.3 Materials and Measures** **40**
4.3.1 General Trust in Automation – Jian et al. (2000) 40
4.3.2 Situational Trust Scale – Automated Driving 40
4.3.3 Driving Tasks 41

4.3.4 Additional measures 42
**4.4 Procedure 43**
**CHAPTER 5. Results 45**
**5.1 RQ1: Is situational trust a separable construct from general trust in automation? 45**
5.1.1 Low Automation 45
5.1.2 High Automation 47
5.1.3 Summary 49
**5.2 RQ2: Is the STS-AD internally reliable? RQ2a: Are there items that can be removed to improve the internal reliability? 49**
5.2.1 Low Automation 50
5.2.2 High Automation 50
5.2.3 Summary 50
**5.3 RQ3: What is the factorial structure of the scale? 50**
5.3.1 Low Automation 50
5.3.2 High Automation 51
5.3.3 Summary 52
**5.4 RQ4: Does automation level change situational trust? 52**
**CHAPTER 6. Discussion 53**
**6.1 Summary of Study 1 - Online Validation 53**
**6.2 Summary of Study 2 – Simulator Validation 54**
6.2.1 The STS-AD after Low Automated Driving Experience 55
**6.3 Theoretical Implications 58**
**6.4 Practical Implications 58**
**6.5 Limitations 59**
**6.6 Future Research 60**
6.6.1 Improving the STS-AD 60
6.6.2 Measuring Situational Trust in Additional Contexts 61
6.6.3 Nuanced Approach to Trust in Automation Measurement 61
6.6.4 Modernized Definition of Automation 61
**CHAPTER 7. Conclusion 63**
**APPENDIX A. Comparison of English and German Speaking Samples by Video 64**
**APPENDIX B. Comparison of STS-AD Scores across Videos within Condition 65**
**APPENDIX C. Demographics Questionnaire for Simulator Validation Study 68**
**APPENDIX D. Participant Instructions for the Driving Simulator Validation 69**
**APPENDIX E. Situation Awareness Rating Technique (SART) 73**
**APPENDIX F. Trust in Automation Scale for Initial Validation 78**
**APPENDIX G. Trust in Automation Scale for Simulator Validation 80**
**APPENDIX H. NASA TLX Definitions 81**
**APPENDIX I. Simulator Validation Debrief Form 82**
**APPENDIX J. How to Use the Situational Trust Scale for Automated Driving (STS-AD) 84**
**APPENDIX K. How to Develop a Situational Trust Scale 85**
**APPENDIX L. Descriptive Statistics for the STS-AD from the Simulator Evaluation 88**
**REFERENCES 90**

# LIST OF TABLES

Table 1. Example items from Trust in Automation Scale (Jian et al., 2000). ..................15
Table 2. Example items from Körber (2019). Responses range from “Strongly disagree (1)” to “Strongly agree (5)”. Higher levels of trust, as measured by this scale, lead to higher levels of reliance; whereas, lower levels of trust lead to lower levels of reliance (Körber, 2019). ..........................................................................................................16
Table 3. Example items from TOAST, Wojton et. al. (2019). TOAST is comprised of nine items that participants respond to with level of agreement ranging from “Strongly agree (7)” to “Strongly disagree (1)”.....................................................................16
Table 4. Items in the STS-AD scale and the factors of situational trust to which they are related. ........................................................................................................................25
Table 5. Number and percentage of participants in each condition. ................................26
Table 6. Significant independent t-tests comparing average STS-AD scores across conditions. All other comparisons were not significant. ..................................................28
Table 7. Factor weightings for Trust in Automation scale and STS-AD that resulted from a Varimax rotation. AV is an abbreviation for Automated Vehicle. ................................33
Table 8. Factor loadings for each item of the Situational Trust Scale..............................34
Table 9. Familiarity level of participants with ACC and ALK systems. ..........................38
Table 10. Rotated factor loading matrix for the STS-AD and Jian et al. (2000) trust scale with the low automation scale responses..........................................................................46
Table 11. Summary of factor loadings for the low automation responses. .......................47
Table 12. Rotated factor loading matrix for the STS-AD and Jian et al. (2000) trust scale with the high automation scale responses. .......................................................................48
Table 13. Summary of factor loadings for the high automation responses. ......................49
Table 14. Communalities of the low automation responses computed in the PCA. .........51
Table 15. STS-AD factor loadings for the high automation drive. ..................................52

# LIST OF FIGURES


Figure 1. Levels of automation for automated driving (National Highway Transportation Safety Administration, n.d.). ....................6
Figure 2. Model of Trust in Automation proposed by Lee and See (2004). ....................9
Figure 3. Explanation of trust calibration posed by Lee and See (2004). ....................10
Figure 4. Model of Trust in Automation (Hoff & Bashir, 2015). ....................11
Figure 5. Relationship between dispositional, situational, and learned trust (Hoff & Bashir, 2015). ....................14
Figure 6. Situational trust scores across videos by condition. ....................28
Figure 7. Average scores for each item of the situational trust scale across videos in Condition 1 (near miss experienced in video 1). Dark blue bars indicate positively worded items. Light blue bars indicate negatively worded items. ....................29
Figure 8. Average scores for each item of the situational trust scale across videos in Condition 2 (near miss experienced in video 2). Dark blue bars indicate positively worded items. Light blue bars indicate negatively worded items. ....................30
Figure 9. Average scores for each item of the situational trust scale across videos in Condition 3 (near miss experienced in video 3). Dark blue bars indicate positively worded items. Light blue bars indicate negatively worded items. ....................30
Figure 10. Average scores for each item of the situational trust scale across videos in Condition 4 (near miss experienced in video 4). Dark blue bars indicate positively worded items. Light blue bars indicate negatively worded items. ....................31
Figure 11. Average scores for each item of the situational trust scale across videos in Condition 5 (near miss experienced in video 5). Dark blue bars indicate positively worded items. Light blue bars indicate negatively worded items. ....................31
Figure 12. Average scores for each item of the situational trust scale across videos in the Control condition (no near miss). Dark blue bars indicate positively worded items. Light blue bars indicate negatively worded items. ....................32
Figure 13. Quarter cab MiniSIM ....................39
Figure 14. The "Play" button provided for participants to touch to begin playing Tetris (left). The game play screen with buttons to rotate shapes and reset the screen (right). ....................40
Figure 15. Map of the routes taken for the simulated automated drives. The low automation drive began at point A and ended at point B. The high automation drive began at point B and ended at point A. ....................42

# SUMMARY

Trust in automation is currently operationalized with general measures that are either self-report or behavioral in nature. However, a recent review of the literature suggests that there should be a more specific approach to trust in automation as different types of trust are influenced by different factors (Hoff & Bashir, 2015). This work is the development and validation of a measure of situational trust for the automated driving context: The Situational Trust Scale – Automated Driving (STS-AD).

The first validation study showed that situational trust is a separable construct from general trust in automation and that it can capture a range of responses as seen in the difference between scores after watching a near automation failure video and non-failure videos. The second study aimed to test the STS-AD in a mid-fidelity driving simulator. Participants drove two routes: low automation (automated lane keeping only) high automation (adaptive cruise control with automated lane keeping). The results of the second study provided further support for situational trust as a distinct construct, provided insight into the factorial structure of the scale, and pointed towards a distinction between advanced driver assistance systems (ADAS) and automated driving systems (ADS).

The STS-AD will revolutionize the way that trust in automation is conceptualized and operationalized. This measure opens the door to a more nuanced approach to trust in automation measurement that will inform not only how drivers interact with automated systems; but, can impact how we understand human-automation interaction as a whole.

# CHAPTER 1. INTRODUCTION AND MOTIVATION

## 1.1 Motivation

Trust in automation is most commonly measured as a general construct and with a single time point scale given after an interaction (e.g., Jian, Bisantz, & Drury, 2000; Körber, 2019; Wojton, Lane, & Porter, 2019). There has been some movement toward dynamic measures of trust in automation, primarily using eye tracking or behavioral outcomes (e.g., Hergeth, Lorenz, Vilimek, & Krems, 2016; Rice & Keller, 2009; Sawyer, Seppalt, Mehler, & Reimer, 2017). However, these measures are still developed for general trust in automation. This is in contrast to the most recent review of the trust in automation literature that proposes many types of trust and culminates in system reliance (Hoff & Bashir, 2015). In order to evaluate the empirical validity of this proposed model, we must develop and validate new measures of trust in automation that are specific to different types of trust. This work is the development and initial validation of the first measure of a type of trust in automation, situational trust, in the automated driving context.

## 1.2 Thesis Statement

The development and validation of a situational trust in automation measurement will further trust in automation research significantly. Trust in automation is measured as a general construct currently. This measurement will give insight into how context of use, task dynamics, and intrapersonal factors influence situational trust in automation. Once this measurement is established, it will be the first measurement of a type of trust in automation rather than general trust in automation. This will lay the groundwork for the ability to empirically validate the model proposed by Hoff and Bashir (2015) and determine how the different types of trust interact with each other and develop over different time courses.

## 1.3 Research Questions

The scale will be evaluated with two key factors in mind: construct validity and internal reliability. The following research questions will frame the evaluation of the scale.

RQ1: Is situational trust a separable construct from general trust in automation?

RQ2: Is the scale internally reliable?

RQ2a: Are there items that can be removed to improve the internal reliability?

RQ3: What is the factorial structure of the scale?

RQ4: Does automation level change situational trust?

## 1.4 Summary of Studies

For several years I have been working to develop a new measure of situational trust that takes into account the specificity of the task and context of automated driving. The Situational Trust Scale – Automated Driving (STS-AD) provides the first empirical evidence for the distinction between situational trust and general trust.

The subsequent study was aimed at validating this measure in a driving simulator with two different levels of automation. The use of the scale in a driving simulator study will add further understanding to the stability of the scale across different experimental use cases. This will allow for comparing the reliability and validity metrics between the online and simulator studies, as well as determine if the factorial structure is stable across fidelity levels. Finally, this study will allow for determining if level of automation is a factor that contributes to situational trust in the automated driving context.

# CHAPTER 2. BACKGROUND

## 2.1 Automation

Allocation of task(s) to a machine or system that a human once performed is automation (Parasuraman & Riley, 1997). There are many different ways automation is implemented and it spans many different contexts. For example, a simple calculator is an example of automation as it reallocates the task of computing sums and multiplication to a machine rather than a human. On the other hand, robots used for bomb detonation by the military are also considered automation. The allocation of the bomb detonation task that the military used to rely on humans to do, in this case, has been reallocated to a robot. These two examples highlight the extremes of many types of automation that are possible. There are also, even within a given system, different levels of automation. For example, a vehicle can have automated lane keeping only, requiring the driver to engage with the accelerator and decelerator to drive the vehicle. Alternatively, a vehicle could have the capability of navigating to a final destination. As more functions are allocated to the machine rather than the human, the higher the level of automation is (Endsley & Kaber, 1999; Parasuraman, Sheridan, & Wickens, 2000; The HART Group, 2011).

Automation is intended to reduce the workload of human operators by offloading tasks to machines or systems and lead to reduced human error (Hancock et al., 2013). However, there are also implications for human performance through a reduction in situation awareness, or knowing what is going on in the environment (Hancock et al., 2013). As tasks are removed from human operators and given to machines, humans are known to become complacent, relying too much on the automation, and also known to suffer from the *generation effect*, having a reduced memory of actions completed by the machine rather than by themselves (Hancock et al., 2013). Although the original intent of automation was to reduce inputs and decisions required from humans, it has added complexity to systems through newfound effects of automation on human performance, including in automated vehicles.

## 2.2 Automated Vehicles

Automated systems have become more prevalent in vehicles in recent years. Implementation of this technology in the driving context was rooted in increasing safety. As much as 94 percent of serious vehicle accidents are due to human error stemming from impaired driving, distracted driving, and not following roadway laws (US Department of Transportation, 2018). Increasing the level of automation in vehicles could reduce the chances of these accidents due to human error in the future. To simplify and standardize the discussion of levels of automation for automated vehicles specifically, the National Highway Transportation Safety Administration (NHTSA) adopted levels proposed by the Society of Automotive Engineers (SAE) seen in (SAE International, 2014). As is common in automation level taxonomies, the lower levels of automation require more human input and decision making than the higher levels of automation (The HART Group, 2011; Wickens, 2018; Wickens, Li, Santamaria, Sebok, & Sarter, 2010).

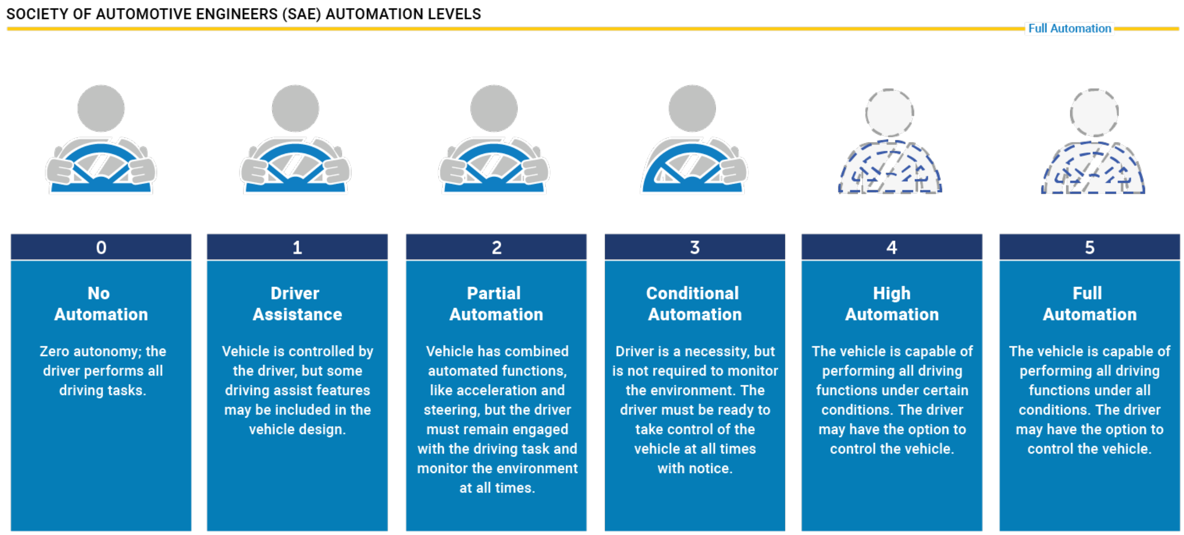


**Figure 1. Levels of automation for automated driving (National Highway Transportation Safety Administration, n.d.).**

## 2.3 Trust in Automation

Trust in automation is crucial for ensuring that automation is used appropriately and that potential benefits are realized (Parasuraman & Riley, 1997). Trust research began in social psychology describing interpersonal trust. Many models and definitions of trust resulted from this research. Mayer's (1995) definition of trust, however, is highly cited in both the human-human trust literature as well as the human-automation trust literature, "willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party," (Mayer, Davis, & Schoorman, 1995, p. 712). This definition highlights three key aspects of trust: vulnerability, expectations, and lack of control or monitoring. Vulnerability, or risk, is required in order for trust to be necessary (Li, Holthausen, Stuck, & Walker, 2019; Stuck, Tomlinson, & Walker, n.d.) One must accept the risk in order to trust the other party.

Lastly, one may or may not be able to monitor or control the other party while the task is being completed, again requiring willingness to be vulnerable to the other party while the task is being completed. This definition encompasses the components of human-human trust; however, it has also been highly referenced in the human automation trust literature.

### *2.3.1 Definition*

Although Mayer's (1995) definition is highly applicable to trust in automation, several definitions have been developed for the trust in automation context specifically. Perhaps the most popular of these is from Lee and See (2004), "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" (p. 54). This definition highlights the task-oriented nature of using automation and emphasizes the necessity of undertaking vulnerability, or risk, in order to use the automation. Both of these aspects are also highlighted by Mayer's definition of interpersonal trust (Mayer et al., 1995). Another definition highlights three areas necessary for trust development, "trust is built on the possibility to observe the system's behavior (performance), understand the intended use of the system (purpose), as well as understand how it makes decisions (process)" (Ekman, Johansson, & Sochor, 2018, p. 96). This definition by Ekman and colleagues overlaps with Mayer's (1995) definition on performance. However, Ekman (2018) does not highlight the risk associated with automation use which we know to be very important (Li et al., 2019; Stuck et al., n.d.)

While there are specific definitions for human-automation trust, Mayer's (1995) definition still holds true in this context. In automated driving, the driver of the vehicle must be willing to accept some risk simply by deciding to use the automated systems in the vehicle. Next, the driver must have some expectations of vehicle performance in the given driving environment. Finally, the driver accepts that they may or may not be able to control or monitor the automation itself. The components of Mayer's (1995) definition are clearly applicable to automation broadly and to the automated driving environment.

## 2.4 Models of Trust in Automation

While definitions provide initial insight into how trust in automation is characterized in the literature, models provide further understanding of the factors that influence trust in automation and the relationships between them. Rather than an exhaustive review of the models of trust in automation, this section is focused on two models that have made a great impact in the research since their publication. Specifically, these models have impacted trust in automated driving research and have led to the development of interventions to support trust in automation.

### *2.4.1 Lee and See (2004)*

Lee and See's (2004) model identifies several nested feedback loops that influence the appropriateness of trust in automation. This model includes context of use including social factors such as organizational and cultural influences. Trust is not a factor included in this model directly. Instead, appropriateness of trust is included as a moderator for the relationship between trust evolution and intention formation as well as the relationship between automation and display.

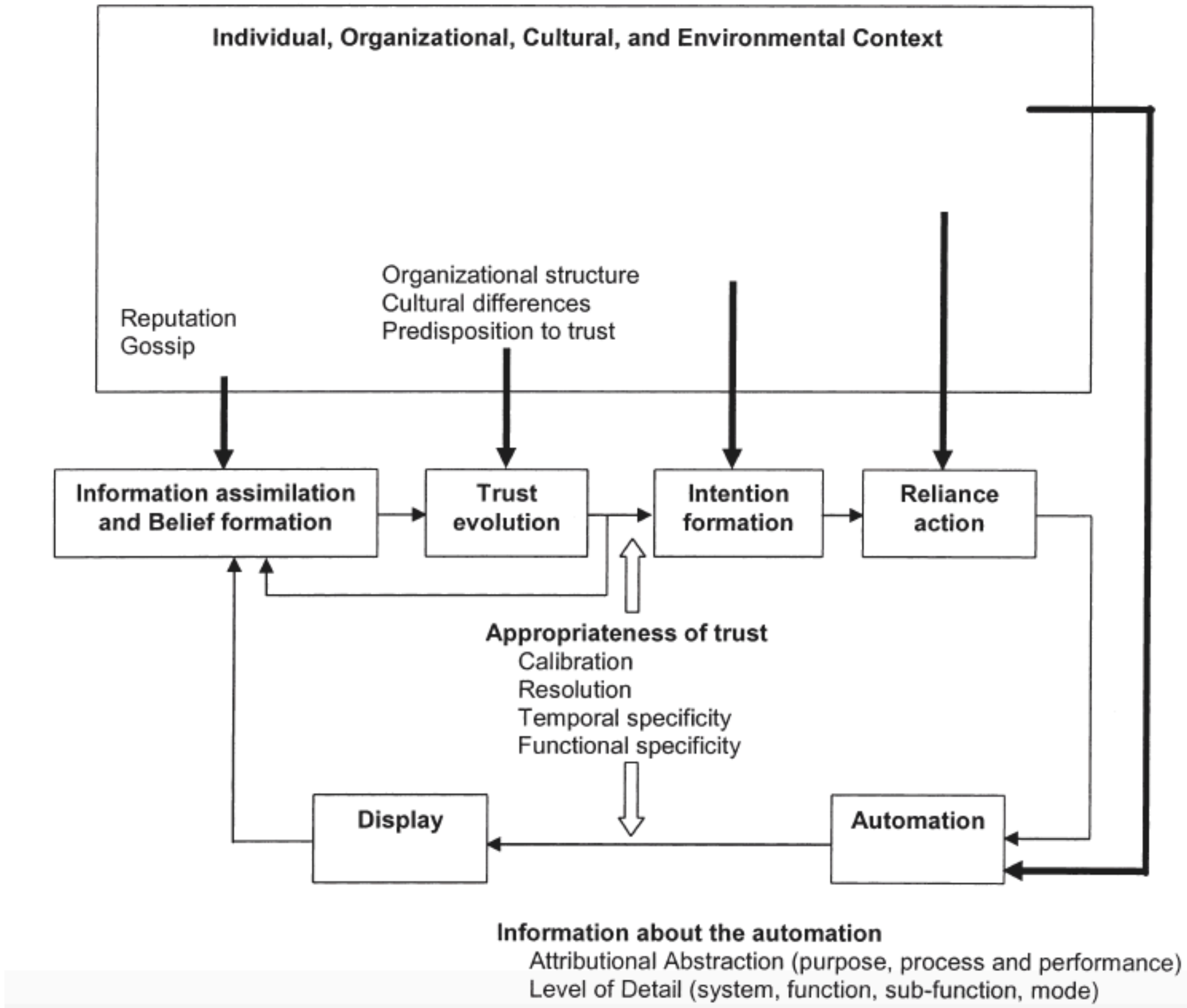


**Figure 2. Model of Trust in Automation proposed by Lee and See (2004).**

The Appropriateness of Trust factor moderates both the relationship between trust evolution and intention formation and the relationship between automation and display. Although this was not the first discussion of what has become known as trust calibration, the authors provide a graphic, seen in Figure 3, that shows the potential for "over and under trust" that fueled research in understanding this construct and also understanding how automation displays influence trust calibration. As seen in Figure 3, trust calibration is defined as a matching of system capabilities with trust in the system (Lee & See, 2004). As stated by Parasuraman and Riley (1997), over-trust leads to misuse and under trust or distrust leads to disuse, neither of which maximize the potential benefits of automation. The Lee and See paper reignited the discussion of trust calibration leading to an area of research discussing how best to present automation uncertainty, also known as automation reliability, to users of automation (e.g., (Beller, Heesen, & Vollrath, 2013; Helldin, Falkman, Riveiro, & Davidsson, 2013; Noah, 2018; Noah, Gable, Chen, Singh, & Walker, 2017; Noah, Gable, & Walker, 2016; Noah & Walker, 2017; Seppelt & Lee, 2007).

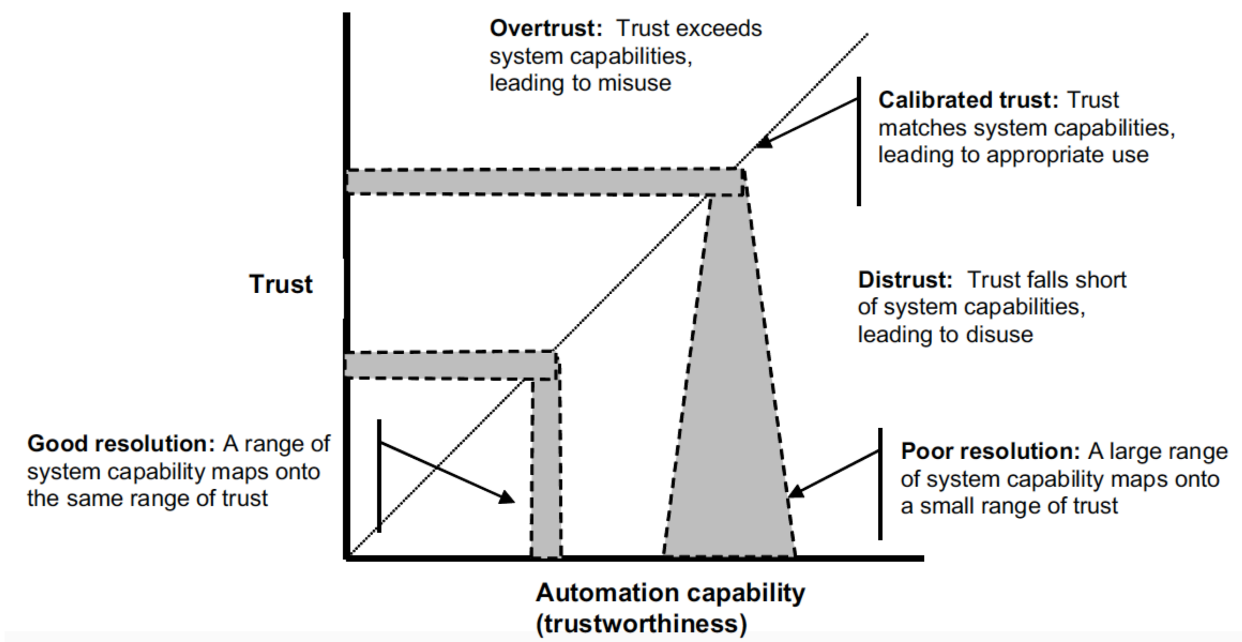


**Figure 3. Explanation of trust calibration posed by Lee and See (2004).**

### *2.4.2 Hoff and Bashir (2015)*

A more recent model of trust in automation was developed from a model of interpersonal trust originally proposed by (Marsh & Dibben, 2005) as seen in Figure 4. In this model, the authors make a distinction between what happens prior to using a system and while using a system (Hoff & Bashir, 2015). Prior to use, the authors, similar to Lee and See (2004), describe an initial reliance strategy. This is a strategy determined by the would-be-user of the automation that describes the intended use of the system and reliance. From there, the authors describe three types of trust in automation that parallel the interpersonal model: dispositional trust, situational trust, and learned trust.

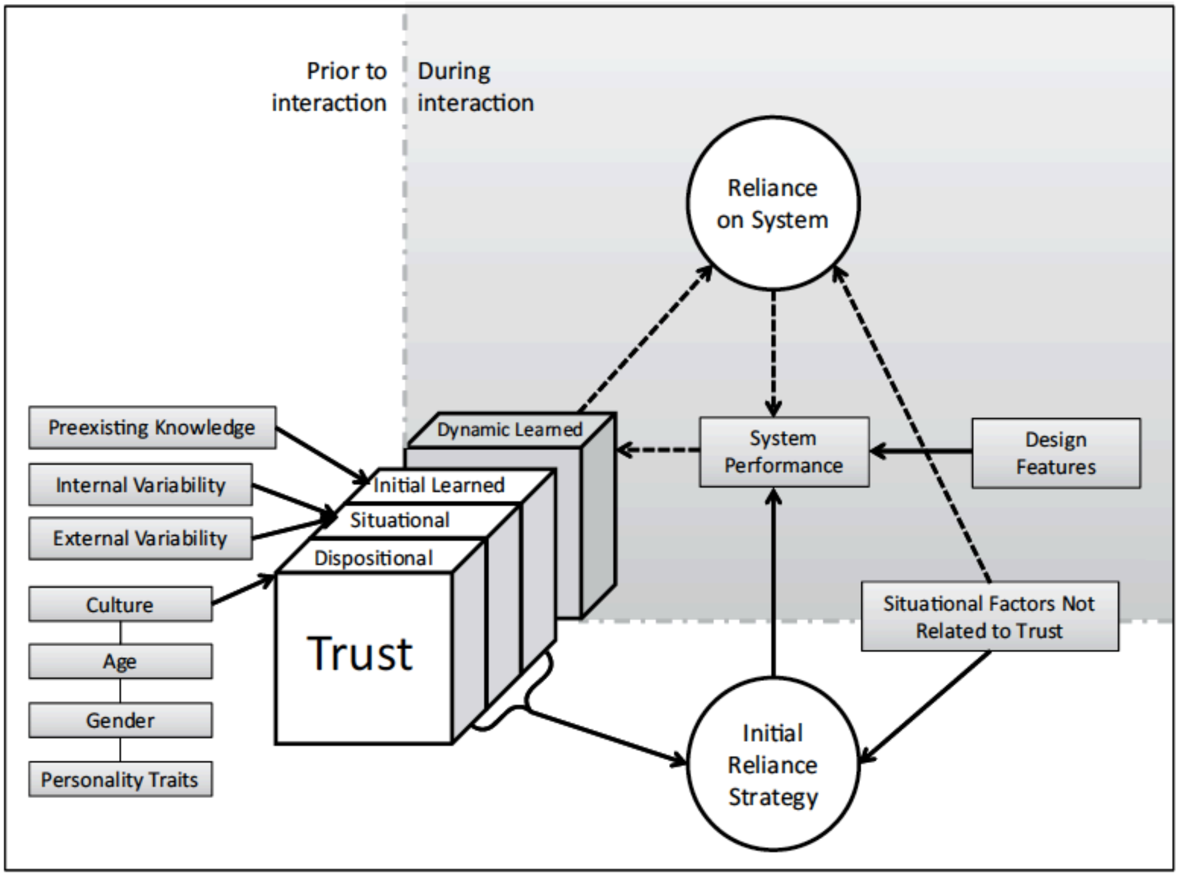


**Figure 4. Model of Trust in Automation (Hoff & Bashir, 2015).**

Dispositional trust can be thought of as a trait, or something that is generally stable across the adult lifespan. This is also sometimes referred to as propensity to trust (Frazier, Johnson, & Fainshmidt, 2013) and propensity to trust automation (Merritt & Ilgen, 2008; Stuck & Rogers, 2018). While previous work has distinguished between propensity to trust, openness to trusting others generally, dispositional trust, and openness toward trusting a specific other, Hoff and Bashir (2015) describe this part of their model as a general propensity independent of the specific entity being trusted (Merritt & Ilgen, 2008). Some factors identified as impacting dispositional trust are culture, age, gender, and personality. If someone has a high dispositional trust, then they would be expected to be open to trusting automated systems; alternatively, if someone has a low dispositional trust, then they would be unlikely to trust automated systems.

Situational trust, on the other hand, is highly dynamic. Hoff and Bashir separate situational trust into two subtypes based on sources of variability: external and internal. External variability of situational trust is caused by risk perceived by the user, the setting in which the automation is being used (e.g., within an organization), the complexity of the system and task, the benefits perceived from using automation, user workload, the difficulty of the task, and the framing of the task. For example, in an SAE Level 3 automated vehicle, there are many automated subsystems, such as adaptive cruise control and automated lane keeping, that work together toward operating the vehicle. This adds a great deal of complexity to the system which, in turn, will increase the external variability of situational trust. Alternatively, in the same Level 3 automated vehicle scenario, the driver may not perceive much risk in using the automation if they only use it on highways, thereby reducing the amount of external variability influencing situational trust. For internal variability, Hoff and Bashir (2015) describe four sources of variance: self-confidence, subject matter expertise, mood, and attentional capacity. If someone perceives themselves as highly competent at completing the task, they will not rely on the automation as much as someone who perceives them self to be less competent at the same task. In the context of attentional capacity, the owner of an automated car may rely more on the automation when they are knowingly impaired, such as when they are highly fatigued. This would reduce the attentional capacity required for the time that they are in the vehicle allowing them to exert less effort. Situational trust, in contrast to dispositional trust, is highly dynamic in that it can change due to a variety of factors in the environment (external variability) and within the individual (internal variability).

Lastly, learned trust is separated into two types: initial learned and dynamic learned. Initial learned trust is knowledge that exists within the individual prior to using a given automated system. This could be from prior experiences with similar systems, a general understanding of how a system functions, reputation of the specific system or manufacturer of the system, and more general expectations of the systems.

Dynamic learned trust, alternatively, is part of a feedback loop that influences reliance on the system itself, which influences system performance (influenced by system design features) and feeds back into dynamic learned trust. The system performance factors identified by Hoff and Bashir (2015) that directly influence dynamic learned trust are: reliability, validity, predictability, dependability, timing of errors, difficulty of errors, type of errors, and usefulness. Design features identified by Hoff and Bashir (2015) that influence system performance are: appearance, ease-of-use, communication style, transparency/feedback, and level of control. Hoff and Bashir propose that dynamic learned trust can change within a given interaction as system performance changes. This is where trust calibration comes into their model of trust in automation. As performance of the system changes during a given interaction, the users trust also changes in line with system performance, ideally. An important distinction between expertise is needed to truly separate situational and initial learned trust. If a given user has expertise within a specific domain, e.g. piloting aircraft in poor weather conditions, then this will influence their situational trust, as they have experience specific to an environmental situation – piloting in poor weather. If a given user has experience with automated systems, this experience would influence initial learned trust as it changes the knowledge that they have about the automation itself prior to interacting with a novel automated system.

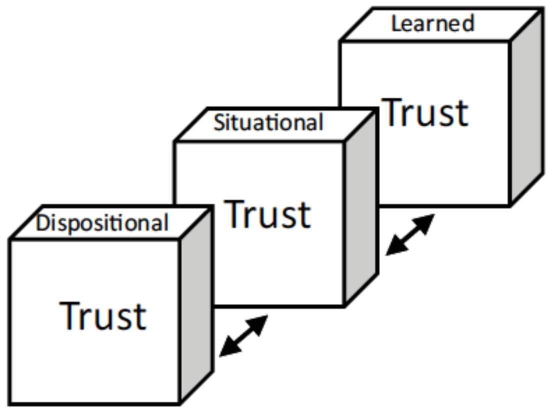


**Figure 5. Relationship between dispositional, situational, and learned trust (Hoff & Bashir, 2015).**

The authors describe the layers of their model, dispositional trust, situational trust, and learned trust as influencing each other as seen in Figure 5. However, no further discussion of exactly how dispositional trust influences situational (and vice versa), or how situational influences learned trust (and vice versa) is provided.

## 2.5 Trust in Automation Measurements

Thus far, all measures of trust in automation have been general – meaning that they are not aimed at measuring a specific type of trust in automation as outlined by Hoff and Bashir (2015). There are two primary types of measures: self-report and behavioral.

### *2.5.1 Self-Report Measures of Trust in Automation*

Trust in automation is most commonly measured through self-report scales or single items. The most frequently used scale in this area is a 12-item assessment of trust and distrust (Jian et al., 2000). Some example items can be seen in Table 1. Participant rate the amount to which they agree or disagree with each statement with anchors, "Strongly disagree" (1) and "Strongly agree" (7).

**Table 1. Example items from Trust in Automation Scale (Jian et al., 2000).**

| Subscale name | Item |
|---|---|
| Trust | The system is reliable. |
| | I am confident in the system. |
| Distrust | The system is deceptive. |
| | The system behaves in an underhanded manner. |

Other trust in automation scales have recently been developed (e.g., Körber, 2019; Wojton, Lane, & Porter, 2019). Both of these scales were developed to measure general trust in automation. Example items for both of these scales can be seen in Table 2 and Table 3. Körber's (2019) scale is unique in that while intended to be used in the context of automation use generally, it was evaluated in the automated driving context. In the process of scale development, the authors used automated driving videos and, finally, automated driving simulations to evaluate the effectiveness of the scale (Körber, 2019; Körber, Baseler, & Bengler, 2018). Alternatively, the Trust of Automated Systems Test (TOAST) was developed in the context of military automated systems.

**Table 2. Example items from Körber (2019). Responses range from "Strongly disagree (1)" to "Strongly agree (5)". Higher levels of trust, as measured by this scale, lead to higher levels of reliance; whereas, lower levels of trust lead to lower levels of reliance (Körber, 2019).**

| Subscale name | Item |
|---|---|
| Familiarity | I already know similar systems. |
| Intention of Developers | The developers are trustworthy. |
| Propensity to Trust | I rather trust a system than I mistrust it. |
| Reliability/competence | The system is capable of interpreting situations correctly. |
| Understanding/predictability | The system state was always clear to me. |
| Trust in Automation | I trust the system. |

**Table 3. Example items from TOAST, Wojton et. al. (2019). TOAST is comprised of nine items that participants respond to with level of agreement ranging from "Strongly agree (7)" to "Strongly disagree (1)".**

| Subscale name | Item |
|---|---|
| Understanding | I understand what the system should do. |
| | I understand the capabilities of the system. |
| Performance | The system helps me achieve my goals. |
| | I feel comfortable relying on information provided by the system. |

Many studies have used scales that are either more specific to the context or experimental manipulations (e.g. Verberne et al., 2012), or unvalidated items (e.g. Waytz, Heafner, & Epley, 2014). Others use several scales together to measure trust and related constructs such as intention to use (Gold, Körber, Hohenberger, Lechner, & Bengler, 2015). These are just a few examples of the ways that trust is measured through self-report items. Despite these new measures, Jian, et al. is still the most highly used scale in the trust in automation literature. Therefore, further analysis in Section 2.5.1.1 will provide evidence to support this as a measure of general trust in automation.

#### 2.5.1.1 Jian et al. (2000) as a General Measure of Trust in Automation

At the time the Jian et al. (2000) scale was developed, the zeitgeist of trust research was just moving towards trust in automation from interpersonal trust. In this time period the focus was on understanding the similarities and differences between interpersonal trust and trust in automation rather than on types of trust as identified by Hoff and Bashir (2015). This can be seen through the methodologies that Jian and colleagues employed to develop the scale.

To develop the Trust in Automation Scale, Jian and colleagues took a tri-phasic approach. First, they completed a word elicitation study. In that study, they asked seven graduate students in linguistics to, "provide written descriptions of their understanding of both trust and distrust with respect to either trust between people, trust in automation, or trust with no specific qualification" (p. 6). This approach, therefore, resulted in words to describe general trust in automation, rather than having words related to a specific type of trust in automation. The research team then followed the word elicitation study with ratings of how related the 138 words were to either trust or distrust when related to general trust, trust in automation, or interpersonal trust. Finally, the authors completed a paired comparisons study. The participants were asked to rate the similarity of two words related to either general trust, interpersonal trust, or human-automation trust. The results of the study were then used to compute factor analyses and cluster analyses which resulted in the final *general* trust in automation scale seen in APPENDIX G. Trust in Automation Scale for Simulator Validation.

The authors describe the resulting scale as, "developed with respect to a non-directed feeling of trust in automated systems, rather than trust in a specific system which the participants had experienced" (p. 32). Therefore, the scale was designed to be generalizable to automated systems broadly. The scale, although appropriate for assessing general trust does not address the constructs identified by Hoff and Bashir (2015) as types of trust in automation: dispositional, learned, and situational.

Dispositional trust is described as a trait; therefore, fairly steady throughout the adult lifespan. However, the Jian et al. (2000) scale has been used repeatedly throughout the literature to show differences in within-subjects experimental manipulations, such as the presence of automation failures, indicating that it measures states, or transient feelings. Researchers have used this scale to measure a state rather than a trait and have shown significant differences between timepoints of administration (e.g. Large, Burnett, Morris, Muthumani, & Matthias, 2018; Noah, 2018). Therefore, it is unlikely that the scale is measuring dispositional trust as it has been used to show state changes due to experimental manipulations within the same participants.

Learned trust refers to either information known prior to interaction with a system, initial learned trust, or knowledge gained while interacting with the system, dynamic learned trust. This scale only has a single item assessing anything related to these constructs, "I am familiar with the system." However, this single item does not assess knowledge truly. It assesses perceived knowledge which could be relevant to these constructs but is not highlighted by Hoff and Bashir.

Finally, situational trust is aimed at being highly contextual in order to understand a person's trust not only towards a specific system but within a specific context of use. The items in the Jian et al. scale do not address any specific's systems task characteristics, benefits of use, or risks of use – all hallmarks of the external variability component of situational trust. Additionally, none of the items are relevant for the internal variability components of situational trust – self-confidence, subject matter expertise, mood, and attentional capacity. Therefore, it is not possible that the Jian et al. scale is measuring situational trust.

While the items cannot be fitted into the Hoff and Bashir (2015) framework, it is possible that researchers could change the instructions and timing of administration of the Jian et al. scale to attempt to measure the constructs defined by Hoff and Bashir. However, we do not truly know how framing changes responses to the items. What we do know is that the Jian et al. scale in its published form does not assess the constructs defined by Hoff and Bashir.

Taken together, this evidence suggests that the Jian et al. scale does not assess any of the constructs put forth by Hoff and Bashir (2015). Instead, it assesses a generalized version of trust in automation that is not specific to any individual, system, or context of use. In order to evaluate the model proposed by Hoff and Bashir, new measures of the types of trust they defined must be developed.

### *2.5.2 Behavioral Measures of Trust in Automation*

Recent work has begun to explore behavioral measures of trust outcomes such as reliance and compliance. The way that each of these measures is conceptualized and measured varies paper-to-paper. Historically, reliance and compliance measures were used to understand trust through behavioral outcomes. Reliance is traditionally defined as, "what the operator does when the automation diagnoses noise in the world," in other words, false alarms (Dixon, Wickens, & McCarley, 2007, p. 564). In contrast, compliance is, "what the operator typically does when the automation diagnoses a signal in the world," or the automation makes a correct target identification (Dixon et al., 2007,

p. 564). Reliance and compliance are often measured with agreement between the operator and automation (e.g., Chancey, Bliss, Yamani, & Handley, 2017; Geels-Blair, Rice, & Schwark, 2013; Rice, 2009; Rice & Geels, 2010; Rice & McCarley, 2011). In automated driving studies measuring reliance and compliance, speed and acceleration are common operationalizations of these constructs (e.g., Cotte, Meyer, & Coughlin, 2001; Yamada & Kuchar, 2006). Furthermore, research is moving towards the use of eye tracking as a measure of trust in automation, specifically in driving as it is dynamic and non-invasive (e.g., Hergeth, Lorenz, Vilimek, & Krems, 2016; Sawyer, Seppalt, Mehler, & Reimer, 2017).

While behavioral measurement is beginning to gain some traction for current and ongoing research, the primary methodology for measuring trust in automation is through post-interaction scales. The gold standard measurement for trust in automation currently is the scale designed by Jian et al. (2000). Future measurements developed in this area should be related to and contrasted with this scale to provide insight into how the novel measurement compares.

# CHAPTER 3. DEVELOPMENT AND INITIAL ONLINE VALIDATION OF THE SITUATIONAL TRUST SCALE - AUTOMATED DRIVING (STS-AD)

This work was completed in close collaboration with my colleague, Philipp Wintersberger. He and I have worked together throughout our time in graduate school and this scale is the culmination of our multi-year collaboration.

## 3.1 Background and Motivation

As aforementioned, trust in automation is currently measured as general trust towards a system. To date, no one has developed a measure of a specific type of trust in automation that allows for a more nuanced perspective.

Hoff and Bashir's (2015) conceptual model was the first definition in the literature of situational trust in automation. Situational trust highlights the impact of contextual differences on trust development as well as how much trust influences behavioral outcomes. Hoff and Bashir (2015) categorized many factors that they believe could contribute to situational trust in automation into two variability subtypes: external and internal. External factors that influence situational trust are type of system, system complexity, task difficulty, workload, perceived risks, perceived benefits, organizational setting, and framing of task. Internal factors that influence situational trust are self-confidence, subject matter expertise, mood, and attentional capacity. These factors were informed by decision aid research primarily. It is our belief that each context of automation use will reveal its own factors that influence situational trust. For example, an automated decision aid in a power plant will have different factors causing situational trust variability than operating an automated vehicle.

In order to evaluate the empirical validity of Hoff and Bashir's (2015) conceptual model we must first be able to measure all of the constructs that they define. This scale development and evaluation is a step towards measuring situational trust in automation in the automated driving context. This scale will allow us to determine if situational trust is distinct from general trust in automation as typically measured.

## 3.2 Methodology

An online survey study was conducted using a series of wizard-of-oz automated driving videos. There were six between-subjects conditions in which the order of exposure to a near automation failure occurred. In Conditions 1-5, the condition number refers to the video in which a near automation failure occurred. In the control condition, there was no near automation failure experienced.

### *3.2.1 Participants*

A total of 303 participants were included in this study. 157 (51.82%) of participants completed the English version of the study and 146 (48.18%) of participants completed the German version. The English-speaking sample was recruited through the psychology recruitment pool at Georgia Institute of Technology. The German-speaking sample was recruited through word of mouth by posting to email lists and online forums.

The average age of the German sample was 36.36 years old (*SD* = 15.81) and the average age of the English sample was 19.78 years old (*SD* = 2.01). The overall average age of the sample was 28.07 years old (*SD* = 13.81). Men comprised 52.15% of the sample (*n* = 158). The English sample reported having their driver's licenses for 3.66

years on average (*SD* = 1.75) compared to the German sample who reported having their licences for 18.41 years on average (*SD* = 15.44). The overall average years of holding a driver's license was 20.25 years (*SD* = 13.04). Participants reported being raised primarily in North America (48.5%) and Europe (48.5%) with a few participants reporting Asia (2%), Africa (0.3%), and South America (0.3%).

### *3.2.2 Materials*

#### 3.2.2.1 Automated driving videos

Six wizard-of-oz automated driving videos were captured for this study. Each video depicted highly automated driving in an urban area in Germany. The videos were captured by a GoPro Hero5, mounted to the windshield of a BMW X5 with automatic transmission. The driver is not visible in the videos and participants were told that the vehicle was automated. Five out of six videos depicted highly reliable automation to the participants. In each of these videos, participants are presented with very similar stimuli – low traffic, smooth turns around road curvature, the same number of lanes, a constant speed, and the same weather conditions (sunny with minimal clouds). The sixth video depicted a near failure. During this video, the vehicle approaches a crosswalk and does not slow down for a pedestrian, who is a confederate, entering the roadway. The confederate pedestrian jumps back away from the car and the car proceeds along the road.

Each participant was presented five of the six total videos. In Conditions 1-5, the condition number corresponds to the order in which the near miss video was presented to the participants (e.g., in Condition 1, the first video was the near failure video; in

Condition 3, it was the third video). In the control condition, all five videos presented highly reliable automation.

#### 3.2.2.2 Situational Trust Scale – Automated Driving (STS-AD)

To measure situational trust, we designed a six item scale that emphasizes the automated driving context's potential risks and benefits as well as the driver's self-efficacy for operating an automated vehicle, the Situational Trust Scale – Automated Driving (STS-AD). As situational contexts can change rapidly while driving (e.g., entering an urban environment from a highway), our intention was to keep the STS-AD as short as possible so that it could be administered quickly and frequently. A 7-point Likert scale (1 – completely disagree; 7 – completely agree) was provided for responding to each item. A 7-point Likert scale was used to a) conform with the Jian et al. (2000) scale response options and b) because of the psychometric properties of 7-point Likert scales (Nunnally, 1978).

**Table 4. Items in the STS-AD scale and the factors of situational trust to which they are related.**

| Item | Situational Trust Factor | Item Abbreviation |
|---|---|---|
| I trust the automation in this situation. | Type of system, system complexity | Trust |
| I would have performed better than the automation in this situation. (Reverse scored.) | Self-confidence, subject matter expertise | Performance |
| In this situation, the automated vehicle performs good enough for me to engage in other activities (such as reading). | Perceived benefits, workload, task difficulty | NDRT (non-driving related task) |
| The situation was risky. (Reverse scored.) | Perceived risks | Risk |
| The automated vehicle made an unsafe judgement in this situation. (Reverse scored.) | Perceived risks | Judgement |
| The automated vehicle reacted appropriately to the environment. | Perceived risks, perceived benefits | Reaction |

#### 3.2.2.3 Trust in Automation Scale

Participants completed a modified version of the Trust in Automation scale developed by Jian et al. (2000) at the conclusion of the study. One item from the original scale was not included in this evaluation due to its misinterpretation in the German version of the scale. In its place, participants were asked about the safety of the automated vehicle. See APPENDIX F. Trust in Automation Scale for Initial Validation.

### *3.2.3 Procedure*

After completing a consent form, participants viewed five videos. Conditions were randomly assigned to participants by the survey engine. Condition number indicates which video depicted the near failure incident. All other videos were presented in a random order determined by the survey engine for each participant. The number of participants in each condition can be seen in Table 5.

**Table 5. Number and percentage of participants in each condition.**

| Condition | Number of Participants | Percentage |
|---|---|---|
| Condition 1 | 56 | 18.5 |
| Condition 2 | 56 | 18.5 |
| Condition 3 | 54 | 17.8 |
| Condition 4 | 44 | 14.5 |
| Condition 5 | 46 | 15.2 |
| Control | 47 | 15.5 |

After watching each video, participants completed the STS-AD. At the conclusion of the study, participants completed the Jian et al. (2000) Trust in Automation Scale.

## 3.3 Results

In order to calculate the situational trust scores, three items of the STS-AD were reverse scored (performance, judgement, and risk). Following this, the responses to each item were averaged to result in a mean level of agreement with each item. All of these results are exploratory in nature to understand the relationship between the items in the STS-AD scale and its construct validity. The goal of these analyses were to: (a) determine if the scale is able to capture variance due to near automation failures as well as constancy with highly reliable automation and similar environments; (b) determine if situational trust is a distinct construct from general trust; (c) determine the reliability of the scale; (d) determine the factorial structure of the STS-AD.

### *3.3.1 Overview*

There was no significant difference in average STS-AD score between the German speaking sample and the English speaking sample $t(296.161) = 0.195$, $p = .845$. Further, a comparison by condition was completed to determine if there were any significant differences by video. No statistically significant differences were found. Therefore, the results as follows combines the English and German speaking samples. All comparison results can be found in APPENDIX A. Comparison of English and German Speaking Samples by Video.

### *3.3.2 Comparing STS-AD scores across videos*

Overall, Situational Trust was significantly lower after participants experienced the near miss of the pedestrian, as seen in Figure 6. Repeated measures ANOVAs revealed that there were significant differences in STS-AD scores across videos. Paired t-tests showed that there were significant differences between STS-AD scores captured immediately after participants experienced the near miss incident compared to the rest of the videos. The test statistics for the *F* and *t*-tests can be found in APPENDIX B. Comparison of STS-AD Scores across Videos within Condition.

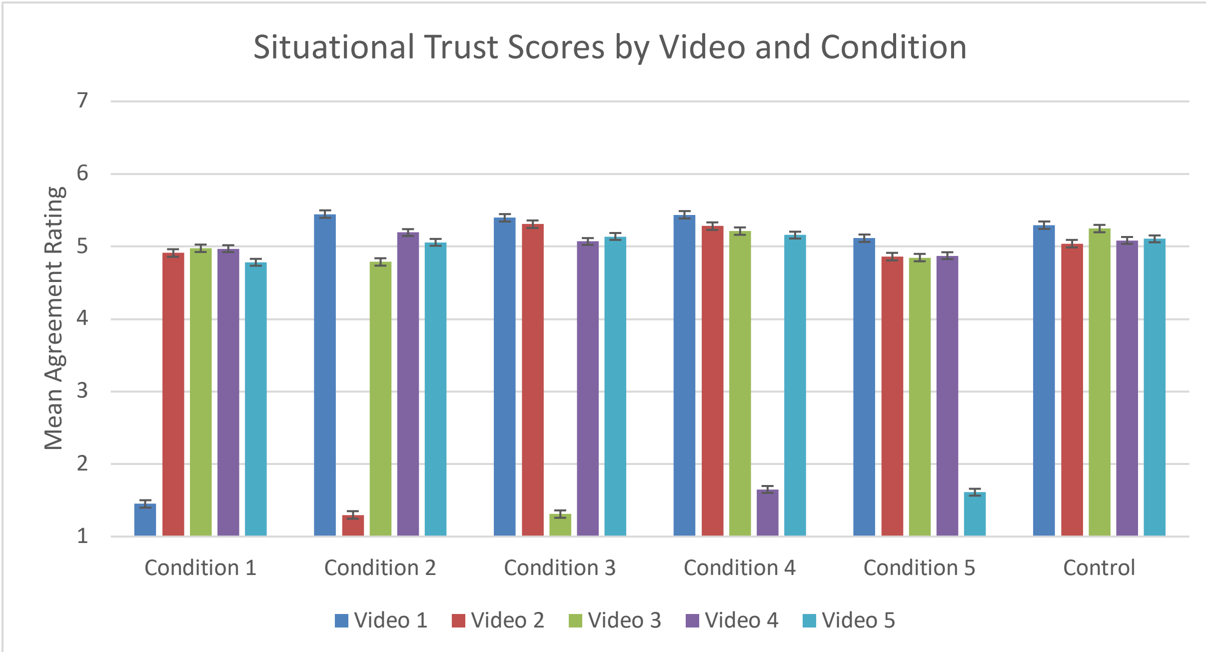


**Figure 6. Situational trust scores across videos by condition.**

There are significant differences in average STS-AD score by condition, $F(5, 297) = 11.098$, $p < .001$. Post-hoc independent sample t-tests were completed with a Bonferonni corrected alpha of 0.0042 to control for family-wise Type 1 error. This analysis revealed that average STS-AD score was significantly higher in the control condition compared to all other conditions. No other comparisons were statistically significant. A summary of the significant results can be seen in Table 6.

**Table 6. Significant independent t-tests comparing average STS-AD scores across conditions. All other comparisons were not significant.**

| Comparison | *t*-value | *df* (equal variances not assumed) | *p*-value |
|---|---|---|---|
| Condition 1 and Control | 5.359 | 92.037 | < .001 |
| Condition 2 and Control | 5.040 | 75.825 | < .001 |
| Condition 3 and Control | 4.455 | 76.909 | < .001 |
| Condition 4 and Control | 3.526 | 84.760 | .001 |
| Condition 5 and Control | 5.520 | 77.847 | < .001 |

Figures 6 through 11 provide insight into how each individual item was rated for each video in each condition. The general pattern of results is low scores across all items for the video in which the failure occurred (if not in the control condition). Otherwise, agreement was rated similarly across items in the highly reliable automation videos. This indicates that the scale is able to capture changes in situational trust due to near automation failures as well as capture near constant levels of situational trust across situations that are highly similar.

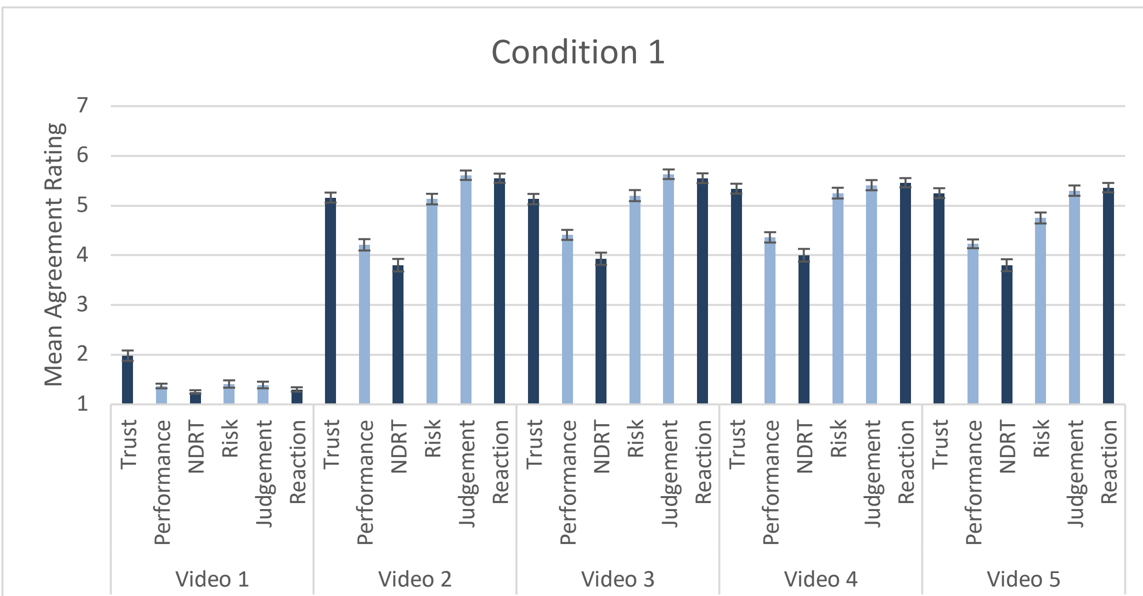


**Figure 7. Average scores for each item of the situational trust scale across videos in Condition 1 (near miss experienced in video 1). Dark blue bars indicate positively worded items. Light blue bars indicate negatively worded items.**

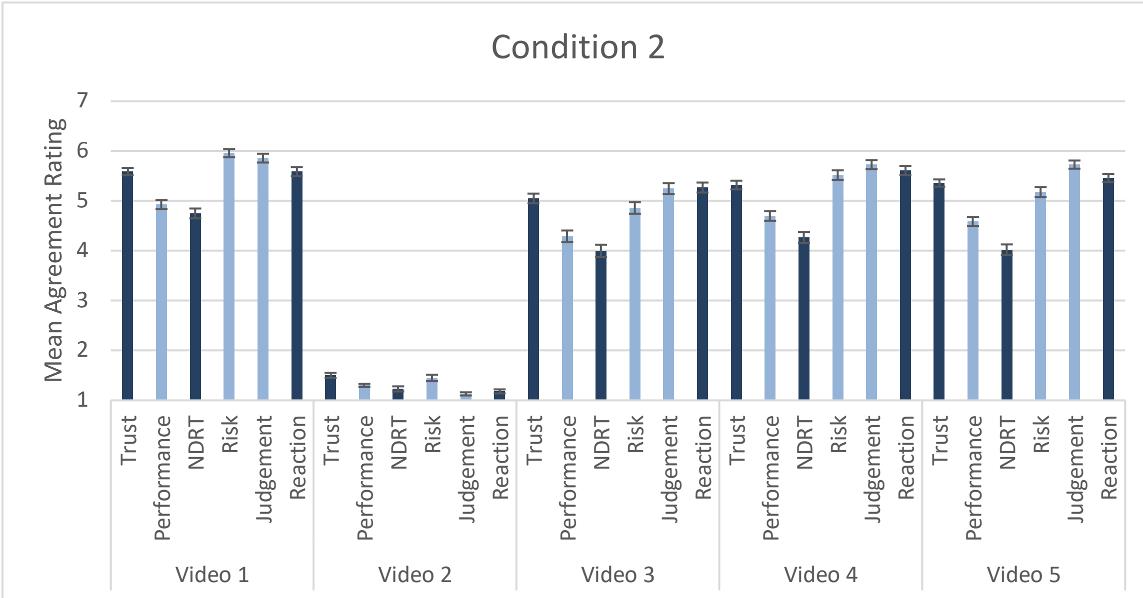


**Figure 8. Average scores for each item of the situational trust scale across videos in Condition 2 (near miss experienced in video 2). Dark blue bars indicate positively worded items. Light blue bars indicate negatively worded items.**

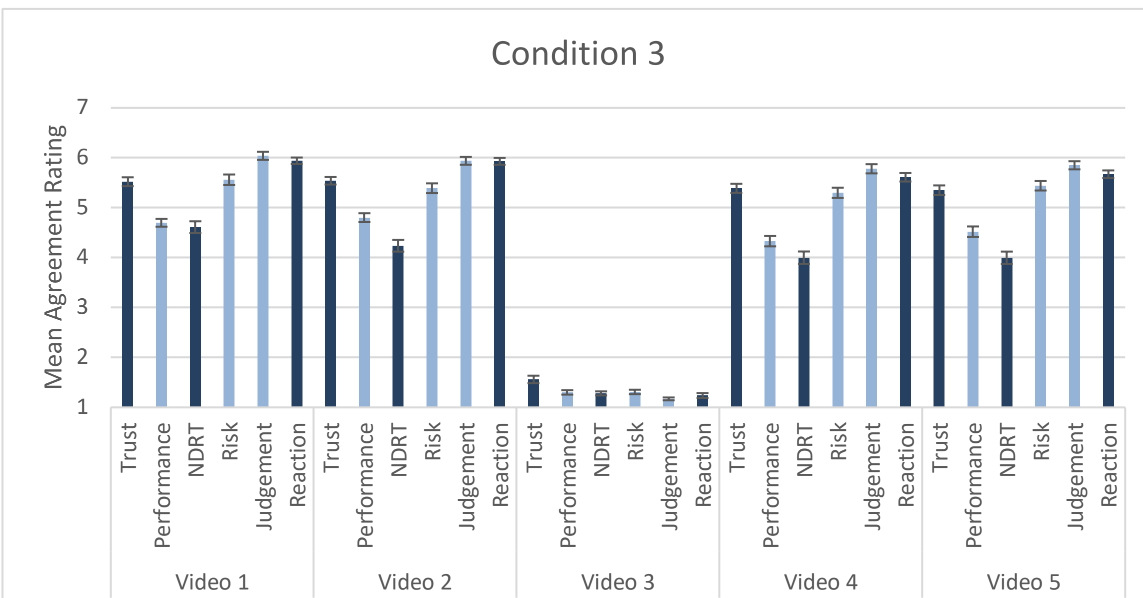


**Figure 9. Average scores for each item of the situational trust scale across videos in Condition 3 (near miss experienced in video 3). Dark blue bars indicate positively worded items. Light blue bars indicate negatively worded items.**

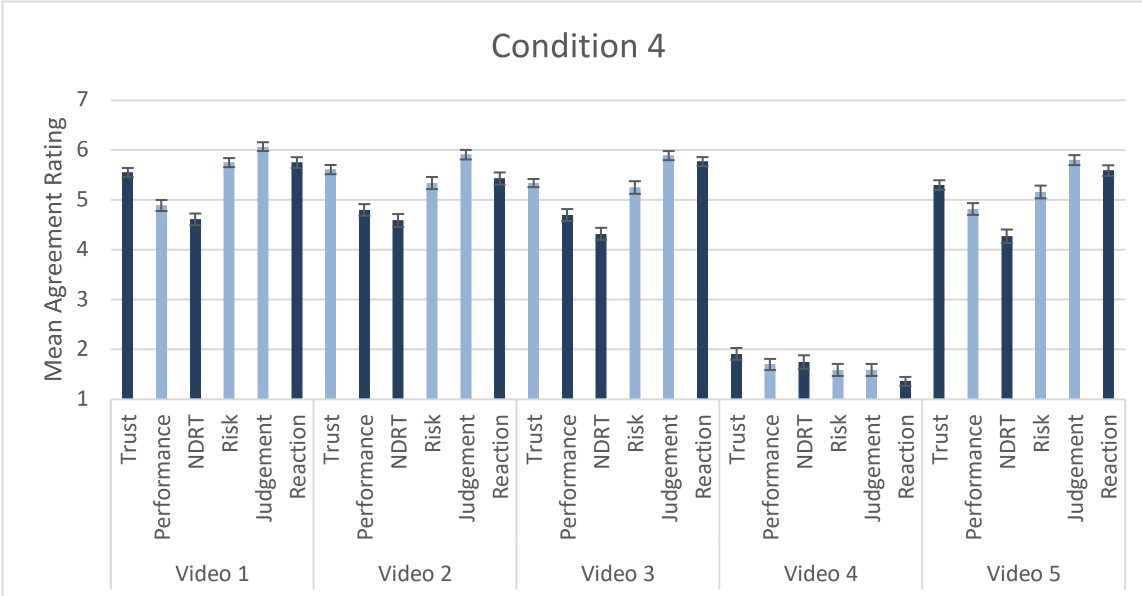


**Figure 10. Average scores for each item of the situational trust scale across videos in Condition 4 (near miss experienced in video 4). Dark blue bars indicate positively worded items. Light blue bars indicate negatively worded items.**

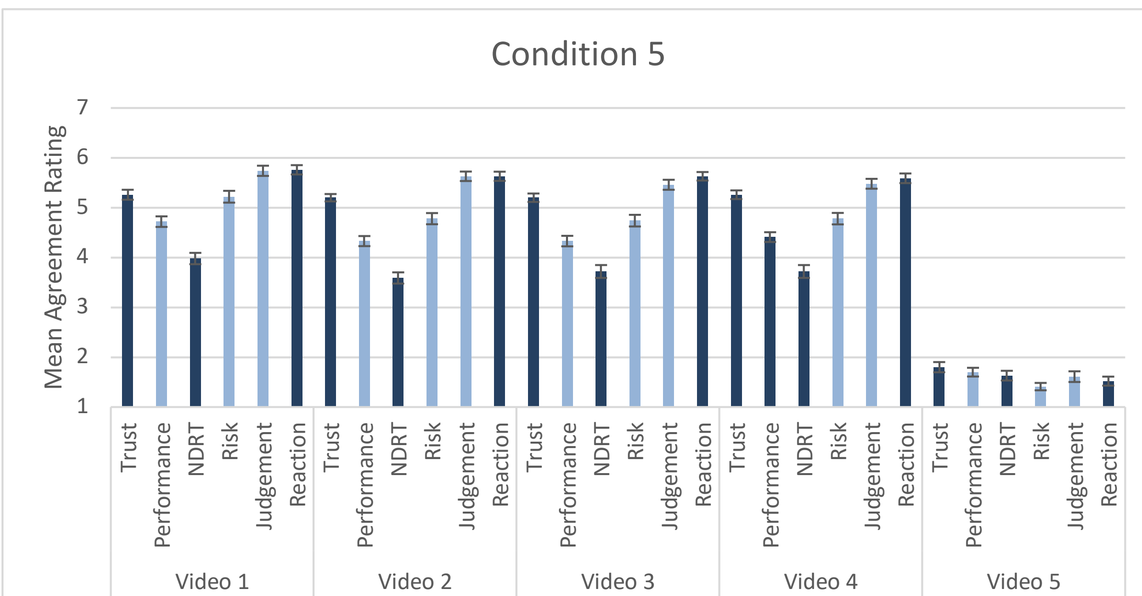


**Figure 11. Average scores for each item of the situational trust scale across videos in Condition 5 (near miss experienced in video 5). Dark blue bars indicate positively worded items. Light blue bars indicate negatively worded items.**

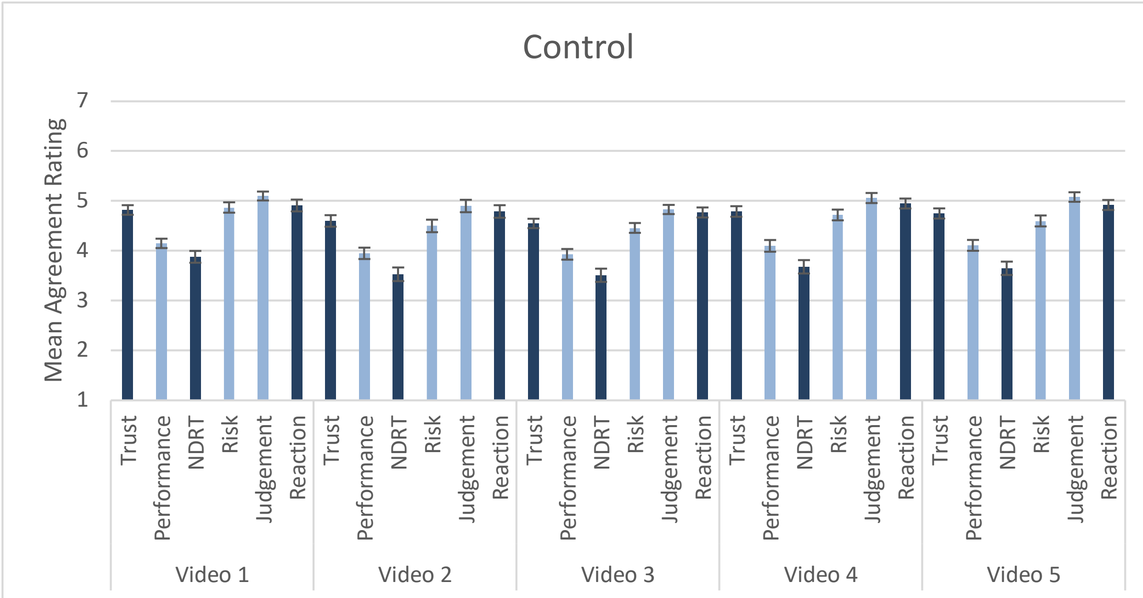


**Figure 12. Average scores for each item of the situational trust scale across videos in the Control condition (no near miss). Dark blue bars indicate positively worded items. Light blue bars indicate negatively worded items.**

### *3.3.3 Construct Validity Evaluation*

To determine how this scale is related to the trust in automation scale, a factor analysis was conducted with the average ratings for each item in both scales. Three factors were extracted, as seen in Table 8. Two factors, Factors 1 and 3, correspond to the trust and distrust subscales, respectively, of the trust in automation scale (Jian et al., 2000). The third factor (Factor 2) corresponds to the STS-AD. Factor weightings are provided in Table 7. This provides further evidence that situational trust is indeed a separate construct from the general trust in automation measured by the trust in automation scale.

**Table 7. Factor weightings for Trust in Automation scale and STS-AD that resulted from a Varimax rotation. AV is an abbreviation for Automated Vehicle.**

| Scale | Item | Factor 1 | Factor 2 | Factor 3 |
|---|---|---|---|---|
| STS-AD | Trust | .341 | .763 | -.034 |
| | Performance | .080 | .550 | -.238 |
| | NDRT | .343 | .528 | .142 |
| | Risk | .119 | .691 | -.213 |
| | Judgement | .034 | .850 | -.218 |
| | Reaction | .057 | .735 | -.208 |
| Jian et al. (2000) | AV is deceptive. | -.325 | -.222 | .731 |
| | The AV behaves in an underhanded manner. | -.094 | -.192 | .804 |
| | I am suspicious of the AV's actions. | -.369 | -.205 | .605 |
| | I am wary of the AV. | -.531 | -.210 | .223 |
| | AV will have harmful outcome. | -.532 | -.196 | .467 |
| | I am confident in AV. | .620 | .173 | -.283 |
| | AV has integrity. | .678 | .081 | -.202 |
| | AV is dependable. | .836 | .122 | -.201 |
| | AV is reliable. | .742 | .115 | -.074 |
| | I can trust the AV. | .884 | .217 | -.180 |
| | I am familiar with this AV. | .373 | .019 | .016 |
| *This item is substituted for the security item that is part of the original scale. | The AV provides safety. | .720 | .185 | -.327 |

### *3.3.4 Inter-item Reliability Analysis*

Cronbach's alpha was used as a measure of inter-item reliability. This analysis was computed after calculating an average for each item in the scale across all videos and conditions. The Cronbach's alpha for the scale is 0.839 indicating that there is a high internal consistency across the six items. Only one item would slightly increase Cronbach's alpha if removed from the scale, "In this situation, the automated vehicle performs good enough for me to engage in other activities (such as reading)." Reliability will be re-evaluated after using the scale in the simulator validation study prior to eliminating any similarly problematic items.

### *3.3.5 Situational Trust Scale Factor Analysis*

To determine if any subscales emerged from this initial evaluation, a factor analysis was computed on the average item responses across videos and conditions. The analysis revealed a single factor indicating that all six items are measuring the same underlying construct. Table 8 shows the factor loadings for each item.

**Table 8. Factor loadings for each item of the Situational Trust Scale.**

| Item | Factor Loading |
|---|---|
| I trust the automation in this situation. | 0.792 |
| I would have performed better than the automation in this situation. | 0.588 |
| In this situation, the automated vehicle performs good enough for me to engage in other activities (such as reading). | 0.532 |
| The situation was risky. | 0.691 |
| The automated vehicle made an unsafe judgement in this situation. | 0.863 |
| The automated vehicle reacted appropriately to the environment. | 0.764 |

## 3.4 Discussion

This measure was specifically developed for measuring situational trust in the context of automated driving. The results of this initial evaluation show empirical support for the distinct construct of situational trust. This provides empirical evidence for the model proposed by Hoff and Bashir (2015), suggesting that further measurement methodologies, specific to types of trust, could be fruitful avenues of future research. This is a significant change in the approach to trust in automation measurement as we move away from general trust measurement towards a more nuanced approach. Measuring types of trust will allow us to understand how the time course of trust development differ across types as well as how automation failures impact types of trust differentially.

STS-AD was found to have a single factor indicating that the items are highly related to a single underlying construct, situational trust. The high reliability of the scale as measured by Cronbach's alpha testing as well as the consistency of scores across similar videos provides strong indications that the scale is providing accurate measurement.

There were several limitations to this study. First, participants were only exposed to automated driving in the form of videos embedded in an online survey. Future evaluations of this scale should move to higher fidelity environments, such as simulation, to ensure that the scale is reliable in a variety of conditions. Although there were consistent results across the similar videos, there was no formal test-retest reliability evaluation for this scale in this initial study as all videos were viewed over a short, single session. Future studies should evaluate test-retest reliability of this scale to ensure consistency.

This scale was specifically developed for the automated driving context which reduces the potential generalizability. However, we believe that this scale could be adapted for other tasks outside of the automated driving domain in the future.

Future work in this area should focus on testing in higher fidelity environments that will provide increased realism for participants. In this high fidelity testing, evaluating the differences between different automation levels would also be an interesting avenue of research. This scale provides initial evidence for the empirical validity of the model proposed by Hoff and Bashir (2015); therefore, future research should explore measuring other types of trust. Once additional measures of types of trust in automation have been developed, evaluating the relationship between them could provide further empirical support for the conceptual model.

### 3.5 Conclusion

This scale provides an initial alternative to measuring general trust. As we move forward with human-automation interaction research we should consider taking a more nuanced approach to measurement rather than simply measuring general trust in automation. This will provide deeper insights into how and why trust in automation changes as well as the time course for change for each type of trust. This approach to trust in automation measurement could also lead to improved trust in automation interventions through more informative displays and improved training.

# CHAPTER 4. METHODS FOR DRIVING SIMULATOR VALIDATION OF THE STS-AD

While the initial evidence for the STS-AD is compelling, to ensure that it is valid in higher fidelity driving environments, we evaluated it in a driving simulator experiment. This experiment had one within-subjects independent variable: automation level. The independent variable had two levels: low and high automation, corresponding to SAE Levels 1 and 3, respectively. In the low automation drive participants only used automated lane keeping (ALK) while operating (i.e. controlling speed and supervising) the simulated vehicle. In the second drive, participants used both ALK and adaptive cruise control (ACC) while operating (i.e. supervising) the simulated vehicle.

Dependent variables included a general measure of trust in automation (Jian et al., 2000) and situational trust in automation measured by STS-AD. Other constructs were measured during the data collection process for this study; however, they are not the focus of the current investigation.

Throughout each driving experience, participants were encouraged to engage in a secondary, or non-driving related task (NDRT), that exploits the same visuo-spatial mental resources as operating a vehicle (Wickens, Tsang, & Benel, 1979).

## 4.1 Participants

In order to enroll in the study, participants were required to have a valid US driver's license and a minimum of two years of driving experience. Additionally, participants had self-reported normal or corrected-to-normal vision, hearing, and sufficient mobility to operate a vehicle.

Fifty-five participants were recruited from the Georgia Institute of Technology Psychology participant pool. Of the 55 participants who completed the study, 10 were excluded from this analysis due to either technical errors or failure to follow instructions, leaving 45 participants for the analysis.

The sample was comprised of 26 males and 19 females. The average age was 20.16 years old (*SD* = 1.89). Participants had their drivers' licenses for 3.61 years on average (*SD* = 1.55). Participants were also asked to report their level of familiarity with automated lane keeping (ALK) systems and adaptive cruise control (ACC) systems. See Table 9 for a summary of the results. Participants reported that they were more familiar with ALK systems than ACC systems. Only 11.11% reported not being familiar with ALK systems while 26.67% of participants reported that they were not familiar with ACC.

**Table 9. Familiarity level of participants with ACC and ALK systems.**

| | ALK | | ACC | |
|---|---|---|---|---|
| Response | Count | Percentage | Count | Percentage |
| I am not familiar this system. | 5 | 11.11% | 12 | 26.67% |
| I am familiar with this system. | 15 | 33.33% | 11 | 24.44% |
| I have been a passenger in a vehicle with this system. | 14 | 31.11% | 11 | 24.44% |
| I have driven a vehicle with this system. | 9 | 20.00% | 8 | 17.78% |
| I own a vehicle with this system. | 2 | 4.44% | 3 | 6.67% |

## 4.2 Apparatus

### *4.2.1 Driving Simulator*

To simulate an automated driving experience, participants drove a quarter cab, mid-fidelity simulator running MiniSIM version 2.2.1, see Figure 13.

**Figure 13. Quarter cab MiniSIM**

### *4.2.2 Non-Driving Related Task (NDRT)*

The secondary task in this experiment is Tetris. This was chosen because of its visuo-spatial nature; requiring participants to use the same cognitive resources they would use while driving (Wickens et al., 1979). The game was played on a touch screen mounted to the right of the steering wheel in the driving simulator. The game was modified so that the blocks fall at a constant rate rather than the rate increasing after a certain number of lines have been cleared. After each line was cleared, participants were awarded 10 points. Participants were able to rotate the shapes and reset the playing area through buttons on the screen, seen in Figure 14. Participants were instructed to earn as many points as possible when they felt it was safe to engage with the game. To begin play, the participant pressed the start button on the touch screen.

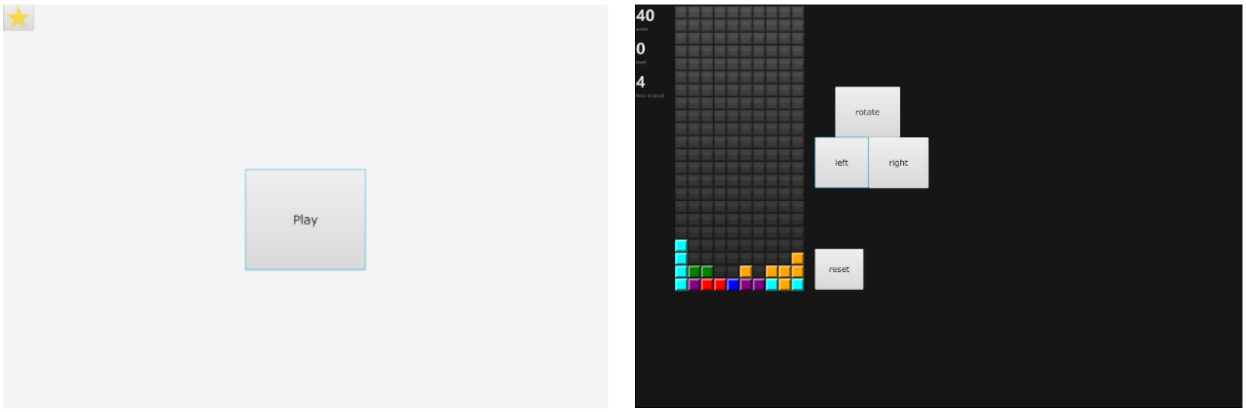


**Figure 14. The "Play" button provided for participants to touch to begin playing Tetris (left). The game play screen with buttons to rotate shapes and reset the screen (right).**

## 4.3 Materials and Measures

### *4.3.1 General Trust in Automation – Jian et al. (2000)*

Participants completed the general trust measure developed by Jian et al. (2000) once at the beginning of the experiment and then twice more after each drive. This scale has 12 items with two subscales: trust and distrust. The full scale is in APPENDIX G. Trust in Automation Scale for Simulator Validation.

### *4.3.2 Situational Trust Scale – Automated Driving*

After each drive participants completed the six-item STS-AD scale described in Section 3.2.2.2.

### *4.3.3 Driving Tasks*

There were two driving tasks in this study. Both drives took place on a rural two lane highway. In the low automation drive, participants used ALK only. In the high automation drive, participants used both ACC and ALK systems. The combination of these systems controlled longitudinal (ACC) and lateral (ALK) position of the vehicle.

Both drives took place on a rural highway with low to moderate traffic and some significant road curvature, seen in Figure 15. Participants were tasked with maintaining a speed of 55 miles per hour (mph). In the low automation drive, participants had to manually control the speed of the vehicle to comply with this instruction. In the high automation drive, the adaptive cruise control system was set to 55 mph. Each drive lasted approximately 20 minutes. The order of the drives was counterbalanced. Figure 15 shows the driving routes. The low automation drive started at point A and ended at point B approximately. The high automation drive started at point B and ended at point A approximately.

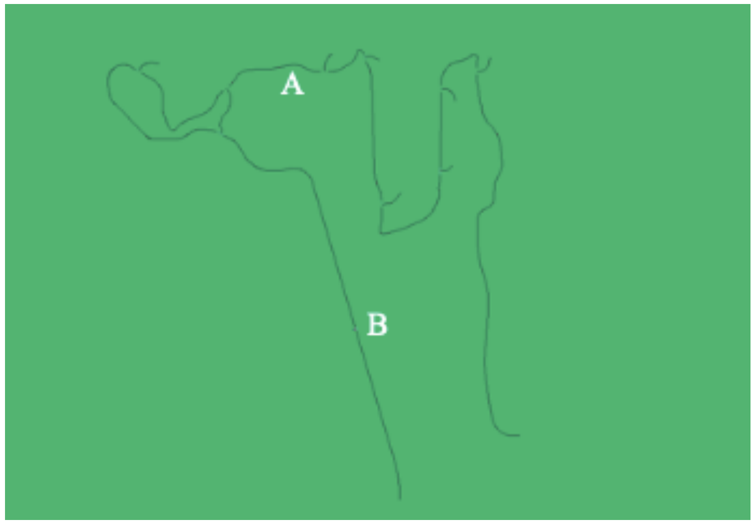


**Figure 15. Map of the routes taken for the simulated automated drives. The low automation drive began at point A and ended at point B. The high automation drive began at point B and ended at point A.**

### *4.3.4 Additional measures*

#### 4.3.4.1 Situation awareness (SA)

Situation awareness (SA) was measured two different ways in this study. First, using the Situation Present Assessment Method (SPAM), participants were asked to respond to probes throughout each driving task (Durso & Dattel, 2004). There were six queries per drive in total. The prompts were presented on the same touch screen as Tetris. At specific locations throughout the driving routes, the experimenter interrupted the Tetris game by presenting a ready prompt. Participants first had to select that they were ready to answer a question. Then, a multiple choice question appeared on screen. Each question asked about past, present, or future events in the drive (two of each category, per drive). The time to respond to the probe and the accuracy of the probe were used as measures of SA.

In addition to this dynamic measurement of SA, participants were also asked to complete the Situation Awareness Rating Technique (SART) after completing each drive (Taylor, 1990). The SART can be found in APPENDIX E. Situation Awareness Rating Technique (SART).

#### 4.3.4.2 Mental workload

Mental workload was measured after each driving experience using the NASA Task Load Index (NASA-TLX) (Hart & Staveland, 1988). The definitions for each NASA-TLX item can be found in APPENDIX H. NASA TLX Definitions.

### 4.4 Procedure

After giving their consent to participate in the experiment, and completing a demographics questionnaire (APPENDIX C. Demographics Questionnaire), participants completed the Georgia Tech Simulator Sickness Screening Protocol to ensure that the driving simulation would not cause them any physical discomfort (Gable & Walker, 2013). Participants were then given instructions for playing the Tetris game and answering the SPAM probes. Full instructions given to participants can be seen in APPENDIX D. Participant Instructions for the Driving Simulator Validation. Then, they completed a trial of these tasks. After being read further instructions about the drives they were about to complete, participants completed their first drive. Following this, they completed SART, Trust in Automation, STS-AD, and NASA-TLX measures. This was followed by the second drive and the same set of measures post-drive. Throughout each drive, participants could engage with Tetris and were asked to respond to SPAM probes.

Finally, participants were debriefed and given credit for their participation. The debrief form can be found in APPENDIX I. Simulator Validation Debrief Form.

# CHAPTER 5. RESULTS

The results will be presented in alignment with the research questions presented in Section 1.3.

## 5.1 RQ1: Is situational trust a separable construct from general trust in automation?

For this analysis, the low automation and high automation conditions were kept separate to ensure that any differences between the responses due to low and high automation were appropriately taken into account.

### *5.1.1 Low Automation*

Combining the items from the Jian et al., (2000) scale with the STS-AD resulted in four factors. Based on an initial Principal Components Analysis (PCA), The factor analysis was completed with a Varimax rotation method. The number of factors was limited to those that could be extracted with eigenvalues greater than 1. The rotated factor loading matrix can be seen in Table 10. A summary of which items loaded onto each factor can be seen in Table 11.

**Table 10. Rotated factor loading matrix for the STS-AD and Jian et al. (2000) trust scale with the low automation scale responses.**

| Scale | Item | Factor | | | |
|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 |
| **STS-AD** | Trust | .455 | -.270 | .092 | .648 |
| | Performance | .162 | -.183 | .171 | -.002 |
| | NDRT | .204 | .103 | .076 | .576 |
| | Risk | -.012 | -.242 | .199 | .379 |
| | Judgement | .150 | -.409 | .661 | .150 |
| | Reaction | .281 | -.122 | .837 | .244 |
| **Jian et al. (2000)** | Deceptive | -.129 | .874 | -.194 | -.214 |
| | Underhanded | -.151 | .732 | -.319 | .007 |
| | Suspicious | -.346 | .892 | -.015 | .163 |
| | Wary | -.397 | .589 | -.018 | -.288 |
| | Harmful | -.257 | .790 | -.153 | -.181 |
| | Confident | .844 | -.257 | .060 | .264 |
| | Security | .754 | -.118 | .209 | .479 |
| | Integrity | .812 | .088 | .413 | .129 |
| | Dependable | .861 | -.345 | .162 | .063 |
| | Reliable | .859 | -.291 | .212 | .164 |
| | Trust | .859 | -.307 | .039 | .125 |
| | Familiar | .578 | -.355 | .129 | .124 |

The STS-AD and the Jian et al. (2000) scale clearly load onto separate factors. While this analysis does not exactly align with the results of initial validation study, which resulted in three factors, it is still clear that the STS-AD is measuring a distinct construct from the Jian et al. scale. The STS-AD scale is spread across factors three and four while the Jian et al. scale loads onto Factors 1 and 2, the trust and distrust subscales respectively.

**Table 11. Summary of factor loadings for the low automation responses.**

| Scale | Item | Factor | | | |
|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 |
| **STS-AD** | Trust | | | | X |
| | Performance | | | X | |
| | NDRT | | | | X |
| | Risk | | | | X |
| | Judgement | | | X | |
| | Reaction | | | X | |
| **Jian et al. (2000)** | Deceptive | | X | | |
| | Underhanded | | X | | |
| | Suspicious | | X | | |
| | Wary | | X | | |
| | Harmful | | X | | |
| | Confident | X | | | |
| | Security | X | | | |
| | Integrity | X | | | |
| | Dependable | X | | | |
| | Reliable | X | | | |
| | Trust | X | | | |
| | Familiar | X | | | |

### *5.1.2 High Automation*

Following the same methodology used for the low automation responses, after an initial PCA, a Varimax rotated factor analysis was completed with the high automation STS-AD responses and the Jian et al. (2000) scale. There were four factors extracted in this analysis, seen in Table 12. Aligned with the low automation results, the Jian et al. items loaded onto Factors 1 and 2 and the STS-AD items loaded onto Factors 3 and 4, seen in

**Table 12. Rotated factor loading matrix for the STS-AD and Jian et al. (2000) trust scale with the high automation scale responses.**

| Scale | Item | Factor | | | |
|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 |
| **STS-AD** | Trust | .406 | -.348 | .354 | .709 |
| | Performance | .120 | -.005 | .479 | .124 |
| | NDRT | .177 | .021 | .221 | .649 |
| | Risk | .212 | -.387 | .387 | .191 |
| | Judgement | .037 | -.417 | .798 | .108 |
| | Reaction | .386 | -.086 | .728 | .195 |
| **Jian et al. (2000)** | Deceptive | -.156 | .848 | -.064 | -.197 |
| | Underhanded | -.068 | .888 | -.157 | .188 |
| | Suspicious | -.315 | .793 | -.030 | .003 |
| | Wary | -.335 | .623 | -.123 | -.351 |
| | Harmful | -.159 | .775 | -.351 | -.332 |
| | Confident | .773 | -.230 | .188 | .461 |
| | Security | .610 | -.169 | .419 | .292 |
| | Integrity | .829 | .098 | .267 | .059 |
| | Dependable | .758 | -.287 | .121 | .048 |
| | Reliable | .819 | -.269 | .299 | .172 |
| | Trust | .810 | -.260 | .264 | .337 |
| | Familiar | .707 | -.225 | -.056 | .125 |

**Table 13. Summary of factor loadings for the high automation responses.**

| Scale | Item | Factor | | | |
|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 |
| **STS-AD** | Trust | | | | X |
| | Performance | | | X | |
| | NDRT | | | | X |
| | Risk | | | X | |
| | Judgement | | | X | |
| | Reaction | | | X | |
| **Jian et al.** | Deceptive | | X | | |
| | Underhanded | | X | | |
| | Suspicious | | X | | |
| | Wary | | X | | |
| | Harmful | | X | | |
| | Confident | X | | | |
| | Security | X | | | |
| | Integrity | X | | | |
| | Dependable | X | | | |
| | Reliable | X | | | |
| | Trust | X | | | |
| | Familiar | X | | | |

### *5.1.3 Summary*

Comparing the two levels of automation for this analysis reveals very similar results. The Jian et al. (2000), scale loads onto Factors 1 and 2 while Factors 3 and 4 are populated with STS-AD items. Because of this, there is empirical support for situational trust being a separate construct from general trust.

## 5.2 RQ2: Is the STS-AD internally reliable? RQ2a: Are there items that can be removed to improve the internal reliability?

To determine if the STS-AD is internally reliable, a Cronbach's alpha analysis was completed on the results from each level of automation.

### *5.2.1 Low Automation*

The overall Cronbach's alpha level of the STS-AD responses after the low automation drive was 0.656. Removing the Performance item would improve the reliability of the scale to 0.686.

### *5.2.2 High Automation*

The overall Cronbach's alpha level of the responses of the STS-AD after the high automation drive was 0.803. There were no items that could be removed to improve the internal reliability of the scale. The high automation responses indicate a high level of reliability (greater than 0.8) as was seen in the initial online validation study.

### *5.2.3 Summary*

These results show some clear differences between using the scale at different levels of automation. The scale is clearly more reliable for use after high automation driving experiences than after low automation driving experiences.

## 5.3 RQ3: What is the factorial structure of the scale?

To answer this question, two exploratory factor analyses were computed, one for each automation level to determine if the scale has a consistent factorial structure if used in the driving simulator environment and across different automation levels. Each of the following factor analyses were conducted after completing initial PCAs. Factors were extracted only if eigenvalues remained greater than one.

### *5.3.1 Low Automation*

When conducting the factor analysis for the low automation results, it was unable to be computed due to communality exceeding one, indicating a Heywood case (Pituch &

Stevens, 2016). Following this initial factor analysis, a second factor analysis was run limiting the number of factors extracted to two, based on the PCA results. This secondary factor analysis resulted in the same Heywood case.

Table 14 shows the communalities that were computed as part of the PCA. Heywood cases often result from misspecification of the factor analysis model, sample size that is too small, or data that are not compliant with the assumptions of factor analysis (Pituch & Stevens, 2016). Regardless of the circumstance(s) that caused the Heywood case, it is clear that the factor analysis results for the entirety of the scale cannot be computed.

**Table 14. Communalities of the low automation responses computed in the PCA.**

| Item | Communality Level |
|---|---|
| Trust | .670 |
| Performance | .288 |
| NDRT | .777 |
| Risk | .290 |
| Judgement | .842 |
| Reaction | .704 |

### *5.3.2 High Automation*

A single factor was extracted from the results of the STS-AD completed after the high automation drive, suggesting a single latent variable, situational trust in automation. The factor loadings are seen in Table 15. The single factor explains 43.64% of the total variance.

**Table 15. STS-AD factor loadings for the high automation drive.**

| Item | Factor Loading |
|---|---|
| Trust | .807 |
| Performance | .470 |
| NDRT | .524 |
| Risk | .598 |
| Judgement | .735 |
| Reaction | .757 |

#### *5.3.3 Summary*

The results of these analyses show that the STS-AD may not be a valid measurement for low levels of automation. However, at the high automation level, the factorial structure was consistent with the results from the online validation, suggesting that the scale does measure a single latent variable, situational trust.

### 5.4 RQ4: Does automation level change situational trust?

The intent of this research question was to understand if automation level can change situational trust. Based on the analyses presented thus far in this chapter, it is clear that the results from the low automation drive lack sufficient reliability, and as indicated by the Heywood case seen in Section 5.3.1, may have some deeper theoretical or structural issues. Therefore, this research question was not explored statistically. Comparing the responses to the scale after the low automation drive to the responses after the high automation drive, from the same participants, indicates there may be some theoretical issues with the definition of automation underlying these results.

# CHAPTER 6. DISCUSSION

As automation, specifically automated vehicles, become more prevalent in everyday life, it is essential to ensure that these systems are being used appropriately to maximize their potential benefits and limit their potential weaknesses. Understanding trust in automation gives insight into perceptions of automated vehicles, how they may be used, and potential interventions to ensure appropriate trust calibration. Thus far, researchers have focused on understanding general trust in automation; however, a recent review points to the necessity of taking a more granular approach to understanding trust in automation through types of trust (Hoff & Bashir, 2015).

Situational trust is influenced by both intrapersonal and external factors, driven by context (Hoff & Bashir, 2015). These factors range from the type of automation being used to the user's self-efficacy and were highlighted without a particular automated system in mind. The STS-AD was developed specifically for the automated driving context to measure situational trust factors that are of particular importance given the task characteristics of operating an automated vehicle.

The development and validation of the STS-AD has established a new approach towards trust in automation measurement. Rather than measuring general trust, as has been the norm for the last two decades, the STS-AD focuses on a new way of approaching trust in automation that is more specific.

## 6.1 Summary of Study 1 - Online Validation

The goal of the online validation study was to understand whether situational trust is a separable construct from general trust and to evaluate the reliability and validity of the STS-AD. The results of this study showed that situational trust is indeed a separable construct from general trust in automation. The results provided evidence that the scale was sensitive to changes in situational trust, internally reliable, and had a single latent variable underlying the six-item scale.

The results of this research established that situational trust could be measured and was distinct from general trust in automation. However, there were limitations to this initial evaluation. Primarily, the study was limited by the fidelity of the driving experience for participants. As this was an online study, participants were only able to experience wizard-of-oz automated driving videos through the survey platforms. The next phase of validation took this into account by moving to a driving simulator as the experimental apparatus.

### 6.2 Summary of Study 2 – Simulator Validation

To address the low fidelity driving experience limitation of the first study, this study aimed to validate the STS-AD in a mid-fidelity driving simulation environment. To determine if the results from the online study would be replicable in this higher fidelity environment, participants completed two drives in the simulator, at low (ALK only) and high (ALK and ACC) automation levels. After each drive participants completed the STS-AD and the trust in automation scale developed by Jian et al. (2000).

The results of the second study showed similar patterns to the first study, with evidence suggesting that the STS-AD is measuring a distinct construct from general trust in automation. However, there were differences seen when comparing the scale evaluations at the different levels of automation.

The responses to the STS-AD after the low automation drive resulted in a lower level of internal reliability and a Heywood case when attempting factor analysis. This leads to the possibility that the STS-AD is not suitable for measuring situational trust in low automation systems. This also points to the possibility that drivers may not perceive the Level 1 automation they experienced in the low automation drive to be the same as the Level 3 automation they experienced in the high automation drive. Perhaps, drivers do not perceive Level 1 to truly be automation at all.

The results of the scale after the high automation drive, on the other hand, align closely with what was seen with the first study. There is a high level of internal reliability and the scale measures single latent variable.

The results of this study provide further evidence for the existence of the situational trust construct and for the reliability and validity of the scale when used after high automation level driving experience.

### *6.2.1 The STS-AD after Low Automated Driving Experience*

The results of the low automation STS-AD were not as expected. There are several possible reasons for this: sample size, wording of the items, and mental models of automation.

#### 6.2.1.1 Contributions of Sample Size

While there are no standards for sample size when doing factor analyses, there are several rules of thumb that could aid in understanding whether this played a role in the results seen here. Traditionally, five participants per variable in the factor analysis was used to ensure an appropriate sample size. In this case with six items, that would indicate a sample size of 30 would be necessary to complete the factor analysis. However, more recent research has shown that significantly higher levels of participants may be necessary. When communalities are relatively high (0.7 or greater) it is possible to use a sample of approximately 100 participants as long as there are three loadings per factor extracted (Fabrigar & Wegener, 2012). However, the communalities observed in this study ranged from 0.288 to 0.842. Therefore, it is suggested, given this pattern that a sample size of 400 or greater may be necessary. While sample size may indeed be a contributing factor, given that the scale results after the high automation drive were in line with the online evaluation, there may be some additional factors contributing to the Heywood case and lower internal reliability seen here.

#### 6.2.1.2 Contributions of Item Wording

Putting the data aside, there is a strong likelihood that the scale was not written to accommodate the mental models that participants have of low automation. Revisiting the STS-AD, in light of this observation, reveals that several items ascribe a level of intelligence that may be inappropriate for low levels of automation. In particular, the Judgement ("The automated vehicle made an unsafe judgement in this situation.") and Reaction ("The automated vehicle reacted appropriately to the environment.") items might impose a level of intelligence that is outside the bounds of an SAE Level 1 automated vehicle. In this driving task, the participant was required to maintain speed and longitudinal position of the vehicle while the automated systems controlled the lateral position within the lane. Given this low level of automation, it is possible that the participants saw the automation as assistive towards the driver rather than truly controlling the vehicle making it difficult, perhaps impossible, for them to know how to respond to the Judgement and Reaction items.

#### 6.2.1.3 Contributions of Mental Models of Automation

There has been a recent movement towards a distinction between advanced driver assistance systems (ADAS) and automated driving systems (ADS) seen in more recent versions of the SAE standard J3016 establishing definitions of automation levels. The automated driving research community is moving towards viewing automation Levels 0-2 as ADAS and Levels 3-5 as ADS. This categorical distinction between levels of automation is entirely different from the way that automation has been discussed in the human factors literature to this point which is as a continuous variable. The results of the STS-AD after the low automation drive point to a distinction in the way that participants in this study viewed the automated driving experiences as well.

#### 6.2.1.4 Recommendations for the STS-AD in Low Automation Situations

Based on the results of Study 2, it is clear that more research is necessary to understand the limitations of the STS-AD to evaluate situational trust in low automation contexts. To that end, future research should explore a broader and larger sample, more representative of drivers in the United States to understand whether this finding is due to this specific sample or whether it holds in one more representative. In addition to further evaluation of the scale in its current form, mental models of participants should be assessed to understand what they perceive as automated versus assistive and what they understand their own tasks as well as the tasks of the vehicle to be. Lastly, the wording of the items may need to be revisited, based on participants mental models of low levels of automation, to assign the appropriate level of intelligence to the system.

## 6.3 Theoretical Implications

This research has provided empirical evidence for the existence of the construct of situational trust in automation, a major leap forward in the study of trust in automation. While theorized in Hoff and Bashir (2015), this is the first instance of evidence supporting the existence of the construct itself. Further, this is the first empirical evidence in support of the Hoff and Bashir model.

The establishment of this construct provides further fodder for continuing to develop and validate measures of the remaining types of trust described by Hoff and Bashir – dispositional trust, initial learned trust, and dynamic learned trust. Further measurement development will provide insights into how these constructs interact with each other and their dynamics over time.

## 6.4 Practical Implications

Further, this work has established a measure of situational trust for automated driving. While further research needs to be conducted to understand the best use of the scale for low levels of automation, the present work clearly shows that this scale is reliable at higher levels of automation. Practitioners interested in understanding how context of use influences automation interaction could use this scale to gain a more nuanced perspective of trust in automation.

While this measurement was developed specifically for the automated driving environment, using the framework set forth here, similar measurements could be developed for any automated system to understand situational trust. This would allow for greater understanding of how different aspects of the human task requirements influence situational trust in automation.

## 6.5 Limitations

There are several limitations to this research that should be considered for future research in this area. First, the STS-AD has a limited number of items. While this was intentional in order to repeatedly administer the scale to participants, future work to expand the number of items could lead to additional diagnosticity of the scale. Several items per factor proposed by Hoff and Bashir could provide insight into which factor(s) are influenced by experimental manipulations.

Given the current state of trust in automation measurement, it is difficult to evaluate construct validity of the STS-AD. As more measurements of types of trust in automation are developed, they should be compared to each other to ensure that they are indeed measuring distinct, albeit related, constructs. Comparison points that are closer together conceptually than the general trust measures will allow for enhanced development of the measurements and improved understanding of the constructs themselves.

A larger, more representative sample in future simulator validation studies could contribute to improved understanding of whether it is appropriate to use the STS-AD for lower levels of automation. Based on the results presented here, current scale is best used in highly automated driving (SAE Levels 3-5). This limited generalizability of the scale is the final limitation. Future work evaluating the parameters laid out in Section 6.2.1 would illuminate the possibility of adjusting and improving the scale to assess situational trust in lower automation contexts.

Finally, there was no assessment of test-retest reliability of the STS-AD in these two studies. However, evaluation of this kind could prove difficult given the sensitivity of this construct. Once a participant has been exposed to a given context of use for a given system, they learn about the from that experience. Presenting the same stimuli a second time and expecting to have the same results may be naïve. However, the internal reliability results presented here for the online validation as well as the high automation simulation indicate that it is likely that the scale would have high test-retest reliability, if evaluated (Cronbach, 1951).

## 6.6 Future Research

### *6.6.1 Improving the STS-AD*

Future research could improve the STS-AD by expanding the number of items to include, at least initially, several per factor proposed by Hoff and Bashir (2015). This expansion of the scale could improve the ability to determine which factor(s) are impacted by particular experimental manipulations.

Although Study 2 provided supportive evidence for the scale in a mid-fidelity simulation environment, additional research in high fidelity driving simulators and on road studies could provide further evidence for the scale's applicability outside of the laboratory environment. This additional data could also point out any differences in results that should be taken into account when using the scale in these settings.

Further evaluation of the scale with a broader range of driving environments, levels of automation, and fidelity of driving experience would aid in understanding the limits of the scale's applicability. To that end, further research aimed at improving this scale should consider using a broad, large sample of participants that is more reflective of the population of United States drivers.

#### *6.6.2 Measuring Situational Trust in Additional Contexts*

The aim of this initial measurement development and evaluation was specifically focused on situational trust in automated driving. However, future work could use the framework provided with this research to develop and evaluate measures of situational trust in any context of automation use. Using the factors outlined by Hoff and Bashir (2015) along with the characteristics of the task environment, items can be developed that mimic those presented here in the STS-AD with context appropriate wording.

#### *6.6.3 Nuanced Approach to Trust in Automation Measurement*

Lastly, this work lays the foundation for a more specific approach towards trust in automation measurement. Rather than continue on with general trust in automation measures, this pushes the field towards exploring types of trust as outlined by Hoff and Bashir (2015). This approach will open up new avenues of research to explore the interaction of these types of trust with each other, level of automation, and context of use. Additional measures of types of trust in automation will also lead to understanding how each construct changes over time to add another layer to our knowledge of trust in automation develops and degrades with automation interactions throughout a lifetime.

#### *6.6.4 Modernized Definition of Automation*

The results of the STS-AD after the low automation drive point to the need to evolve the way that we define and describe automated systems. The Parasuraman and Riley (1997) definition for automation is now over 20 years old. As the technology landscape has evolved in the last two decades it is quite possible that the definition of what is truly perceived to be automation has also changed. According to the Parasuraman et al. (1997) any machine, system, or computer, that takes a task away from a human is automation. By that definition, many things that we use in our day-to-day lives such as toasters and microwaves would be automation. However, in the common vernacular these are referred to as appliances and are fairly ubiquitous, at least in the United States. Updating our definition of automation based upon modern technological advancements would allow us to take these new distinctions into account for experimental manipulations and for measurement development.

# CHAPTER 7. CONCLUSION

The goal of this research was to establish and evaluate a measure of situational trust in automation. Study 1 provided empirical support for the distinct construct of situational trust, showed that a single latent variable underlies the STS-AD, and provided evidence for high internal reliability of the STS-AD. Building upon the findings of Study 1, Study 2 evaluated the STS-AD in a driving simulator environment at low and high automation levels. The results of this study showed that the STS-AD, in its current state, is best used to evaluate situational trust at high levels of automation (SAE Levels 3-5).

Empirical support for the construct of situational trust and the STS-AD will provide a foundation for developing future measures of situational trust in other contexts, as well as expanding the library of trust in automation measures to additional types of trust. This work has shown that measurement of trust in automation should be nuanced, rather than general, to truly understand the factors underlying behavioural outcomes such as automation use.

Future work in this area should consider expanding the size and representativeness of the samples used to evaluate the STS-AD. Additionally, expanding the scale to different levels of automation and higher fidelity driving experiences will provide further understanding of the scale's strengths and limitations. Further, researchers should consider developing a new definition for automation based in modern

# APPENDIX A. COMPARISON OF ENGLISH AND GERMAN SPEAKING SAMPLES BY VIDEO

Equal variances were not assumed for these tests. A Bonferonni corrected alpha of 0.01 was used to determine significance of these tests.

| **Condition** | **Video** | ***t*-value** | ***df*** | ***p*-value** |
|---|---|---|---|---|
| Condition 1 | Video 1 | .610 | 48.423 | .545 |
| | Video 2 | 1.068 | 53.041 | .290 |
| | Video 3 | 2.371 | 53.002 | .021 |
| | Video 4 | .624 | 53.684 | .535 |
| | Video 5 | .235 | 52.738 | .815 |
| Condition 2 | Video 1 | .293 | 53.990 | .771 |
| | Video 2 | .904 | 36.144 | .372 |
| | Video 3 | .864 | 52.722 | .391 |
| | Video 4 | 1.029 | 50.838 | .308 |
| | Video 5 | 1.803 | 53.285 | .077 |
| Condition 3 | Video 1 | 1.811 | 51.117 | .076 |
| | Video 2 | .510 | 51.346 | .612 |
| | Video 3 | .981 | 39.324 | .333 |
| | Video 4 | .341 | 46.119 | .734 |
| | Video 5 | 1.603 | 48.497 | .115 |
| Condition 4 | Video 1 | 0.899 | 28.545 | .376 |
| | Video 2 | 0.699 | 27.930 | .490 |
| | Video 3 | 0.631 | 35.453 | .532 |
| | Video 4 | 2.077 | 28.819 | .047 |
| | Video 5 | 1.487 | 31.214 | .147 |
| Condition 5 | Video 1 | 0.326 | 39.550 | .746 |
| | Video 2 | 1.178 | 43.951 | .245 |
| | Video 3 | 0.078 | 39.463 | .938 |
| | Video 4 | 0.444 | 43.913 | .660 |
| | Video 5 | 0.260 | 43.989 | .796 |
| Control | Video 1 | 0.237 | 44.626 | .814 |
| | Video 2 | .362 | 43.580 | .719 |
| | Video 3 | .035 | 34.234 | .972 |
| | Video 4 | .117 | 43.669 | .908 |
| | Video 5 | .710 | 42.431 | .481 |

# APPENDIX B. COMPARISON OF STS-AD SCORES ACROSS VIDEOS WITHIN CONDITION

The Huynh-Feldt adjustment for sphericity was used for all tests.

| **Condition** | ***F*-value** | ***df*** | ***p*-value** |
|---|---|---|---|
| 1 | 175.405 | 3.835 | < .001 |
| 2 | 204.136 | 3.569 | < .001 |
| 3 | 293.880 | 3.646 | < .001 |
| 4 | 119.118 | 2.958 | < .001 |
| 5 | 110.290 | 3.190 | < .001 |
| Control | 0.923 | 3.315 | .439 |

<u>Condition 1 paired *t*-tests.</u> Significance required to be < .005 due for family-wise correction. Asterisks indicate significant differences.

| **Comparison** | ***t*-value** | ***df*** | ***p*-value** |
|---|---|---|---|
| Video 1 and Video 2 | 19.503 | 55 | < .001* |
| Video 1 and Video 3 | 19.738 | 55 | < .001* |
| Video 1 and Video 4 | 19.640 | 55 | < .001* |
| Video 1 and Video 5 | 19.061 | 55 | < .001* |
| Video 2 and Video 3 | .442 | 55 | .660 |
| Video 2 and Video 4 | .369 | 55 | .713 |
| Video 2 and Video 5 | .718 | 55 | .476 |
| Video 3 and Video 4 | .022 | 55 | .982 |
| Video 3 and Video 5 | 1.083 | 55 | .284 |
| Video 4 and Video 5 | 1.307 | 55 | .197 |

<u>Condition 2 paired *t*-tests</u>

| **Comparison** | ***t*-value** | ***df*** | ***p*-value** |
|---|---|---|---|
| Video 1 and Video 2 | 25.844 | 55 | < .001* |
| Video 1 and Video 3 | 3.614 | 55 | 0.001* |
| Video 1 and Video 4 | 1.913 | 55 | .061 |
| Video 1 and Video 5 | 2.683 | 55 | .010 |
| Video 2 and Video 3 | 17.007 | 55 | < .001* |
| Video 2 and Video 4 | 22.280 | 55 | < .001* |
| Video 2 and Video 5 | 22.409 | 55 | < .001* |
| Video 3 and Video 4 | 2.190 | 55 | .033 |
| Video 3 and Video 5 | 1.312 | 55 | .195 |
| Video 4 and Video 5 | 1.050 | 55 | .298 |

Condition 3 paired *t*-tests

| **Comparison** | ***t*-value** | ***df*** | ***p*-value** |
|---|---|---|---|
| Video 1 and Video 2 | .716 | 53 | .477 |
| Video 1 and Video 3 | 25.349 | 53 | <.001* |
| Video 1 and Video 4 | 2.209 | 53 | .032 |
| Video 1 and Video 5 | 2.254 | 53 | .028 |
| Video 2 and Video 3 | 26.596 | 53 | < .001* |
| Video 2 and Video 4 | 1.874 | 53 | .066 |
| Video 2 and Video 5 | 1.387 | 53 | .171 |
| Video 3 and Video 4 | 22.287 | 53 | < .001* |
| Video 3 and Video 5 | 21.923 | 53 | < .001* |
| Video 4 and Video 5 | .484 | 53 | .631 |

Condition 4 paired *t*-tests

| **Comparison** | ***t*-value** | ***df*** | ***p*-value** |
|---|---|---|---|
| Video 1 and Video 2 | 1.126 | 43 | .266 |
| Video 1 and Video 3 | 1.669 | 43 | .102 |
| Video 1 and Video 4 | 14.500 | 43 | < .001* |
| Video 1 and Video 5 | 1.473 | 43 | .148 |
| Video 2 and Video 3 | .381 | 43 | .705 |
| Video 2 and Video 4 | 13.453 | 43 | < .001* |
| Video 2 and Video 5 | .660 | 43 | .513 |
| Video 3 and Video 4 | 14.404 | 43 | < .001* |
| Video 3 and Video 5 | .333 | 43 | .741 |
| Video 4 and Video 5 | 13.272 | 43 | < .001* |

Condition 5 paired *t*-tests

| **Comparison** | ***t*-value** | ***df*** | ***p*-value** |
|---|---|---|---|
| Video 1 and Video 2 | 1.414 | 45 | .164 |
| Video 1 and Video 3 | 1.432 | 45 | .159 |
| Video 1 and Video 4 | 1.361 | 45 | .180 |
| Video 1 and Video 5 | 12.869 | 45 | < .001* |
| Video 2 and Video 3 | .082 | 45 | .935 |
| Video 2 and Video 4 | .099 | 45 | .921 |
| Video 2 and Video 5 | 14.871 | 45 | < .001* |
| Video 3 and Video 4 | .201 | 45 | .842 |
| Video 3 and Video 5 | 14.291 | 45 | < .001* |
| Video 4 and Video 5 | 13.920 | 45 | < .001* |

Condition 6 – Control paired *t*-tests

| **Comparison** | ***t*-value** | ***df*** | ***p*-value** |
|---|---|---|---|
| Video 1 and Video 2 | 1.262 | 46 | .213 |
| Video 1 and Video 3 | .369 | 46 | .714 |
| Video 1 and Video 4 | 1.468 | 46 | .149 |
| Video 1 and Video 5 | 1.311 | 46 | .196 |
| Video 2 and Video 3 | 1.161 | 46 | .252 |
| Video 2 and Video 4 | .258 | 46 | .798 |
| Video 2 and Video 5 | .321 | 46 | .750 |
| Video 3 and Video 4 | 1.151 | 46 | .256 |
| Video 3 and Video 5 | 1.010 | 46 | .318 |
| Video 4 and Video 5 | .157 | 46 | .876 |

# APPENDIX C. DEMOGRAPHICS QUESTIONNAIRE FOR SIMULATOR VALIDATION STUDY

1. Do you have a driver's license?
    a. Yes (If yes, how many years have you held a license for?)
    b. No (If no, the following message appears: "Sorry, you are not eligible for the study. Please see the researcher.")
2. How many years have you held a license for?
    a. Less than two years (If no, the following message appears: "Sorry, you are not eligible for the study. Please see the researcher.")
    b. Two years or more
3. How many years have you held your license for?
4. On average, how many hours do you drive each week when you're on campus?
5. On average, how many hours do you drive each week when you're not on campus?
6. What is your age?
7. What is your gender?
    a. Male
    b. Female
    c. Other
    d. Choose not to identify
8. What is your primary language?
9. What other languages do you speak?
10. What is your level of familiarity with automated safety features such as automated lane keeping? Automated lane keeping systems automatically steer the vehicle to maintain position within a lane.
    a. I own a vehicle with one or more automated safety features
    b. I have driven a vehicle with one or more automated safety features
    c. I have been a passenger in a vehicle with one or more automated safety features
    d. I am familiar with automated safety features
    e. I have never heard of automated safety features prior to this study
11. What is your level of familiarity with automated safety features such as automated lane keeping? Automated lane keeping systems automatically steer the vehicle to maintain position within a lane.
    a. I own a vehicle with one or more automated safety features
    b. I have driven a vehicle with one or more automated safety features
    c. I have been a passenger in a vehicle with one or more automated safety features
    d. I am familiar with automated safety features
    e. I have never heard of automated safety features prior to this study

# APPENDIX D. PARTICIPANT INSTRUCTIONS FOR THE DRIVING SIMULATOR VALIDATION

**Thanks and Introduction**
First of all, thank you for your participation in this study. We are members of Sonification Lab in school of psychology.

**Purpose of Experiment**
This research is investigating how situation awareness changes as automation increases in vehicles.

**Procedure**

<u>Consent</u>
The consent form presented to you is to inform you of the content of this experiment. Please read through it, and ask any questions you have before you sign it. During the experiment, please let us know if you have questions, concerns, discomforts, or would like to withdraw from the experiment. You can do so without penalty.

<u>General Instructions</u>
Before this experiment, we will ask you to complete a simulator sickness screening first. This is to ensure that you do not encounter any motion sickness during the experiment. Then you will be asked to complete a set of questionnaires. Next, you will complete the first of two driving scenarios which will be followed by another set of questionnaires. The second drive will follow with a final set of questionnaires. The session should last no longer than an hour and a half, and the experimenter(s) will help you throughout the session.

<u>Sim Sickness Screening</u>
To make sure the driving simulator will not cause you any physical discomfort, we will conduct a screening procedure. This procedure includes a pre-drive survey, a short drive, and a post-drive survey. If for any reason, you feel sick during the procedure, this session will end and you will receive full credit for your time here.

<u>Questionnaires and Driving Tasks</u>
We will be collecting data during two separate driving tasks. Both driving courses will be about 20 minutes in length. Each drive will have its own set of instructions that the experimenters will go over with you before the drive. We will be collecting a number of measurements, such as eye movement, driving performance, situation awareness, and workload. The driving scenario will be similar for both of the driving tasks.

<u>Debrief</u>
Once the final set of questionnaires is completed after the second drive, you will be debriefed on the experiment and released. We will then assign you credit for your participation.

**<u>Driving Task 1:</u>**

<u>The Drive</u>
The driving course will last about 20 minutes. For this drive, we would like you to maintain a speed of 55 miles per hour for the duration of the drive to the best of your ability.

- For participants in the **low automation condition (ALK only):**
    - During this drive you will be using automated lane keeping. Automated lane keeping systems keep the vehicle in the center of the lane so that you do not have to steer. Putting your hands on the steering wheel and turning it even a small amount can turn off the automated lane keeping system. To avoid doing so inadvertently, please keep your hands off of the steering wheel unless you feel that you must take control of the vehicle to avoid an accident.
    - You will know that the automated lane keeping system is on if the green, nearly parallel lines are present in the dashboard. If this display is not present, the automated lane keeping system is not on.

- For participants in the **high automation condition (ALK and ACC):**
    - During this drive you will be using automated lane keeping and adaptive cruise control. Automated lane keeping systems keep the vehicle in the center of the lane so that you do not have to steer. Adaptive cruise control systems maintain a preset speed and adjust based on vehicles directly in front of you so that you do not have to accelerate or decelerate. Putting your hands on the steering wheel and turning it even a small amount can turn off the automated lane keeping system. Similarly, if you press either the gas or brake pedal, the adaptive cruise control system will turn off. To avoid doing so inadvertently, please keep your hands off of the steering wheel and your foot off the pedals unless you feel that you must take control of the vehicle to avoid an accident.
    - You will know that the automated lane keeping system is on if the green, nearly parallel lines are present in the dashboard. Similarly, you will know if the adaptive cruise control system is on if a car icon is present in the dashboard. If these displays are not present, the automated lane keeping system is not on.

During the drives you will have the option to engage in a game on the side tablet screen, Tetris. A screen with a "Play" button will be presented and will remain on the screen unless you decide it is safe to play the game. If you press the "Play" button, the game will begin. In Tetris, different shapes fall from the top of the screen and your goal is to completely fill a row. Once you fill a row, it disappears and you are awarded points. If you run out of space, then the game is over and you can reset it for a new round to start. The control buttons will be on the right side of the screen. We will be recording your score. The control buttons will be on the right side of the screen. Periodically, you will be presented questions on the tablet screen. You will know a question is being presented when the “Ready” button appears on the tablet. If you are ready to answer the question, you will simply press the “Ready” button and answer the question on the tablet. After you answer the question, the “Play” screen for Tetris will reappear. Do you have any questions?

Please complete the drive as safely as possible.

Completion

Upon completion, we will stop the driving scenario and present you with questionnaires to complete prior to beginning the second drive.

**Driving Task 2:**

The Drive

The driving course is similar to the last drive you just completed. For this drive, we would again like you to maintain a speed of 55 miles per hour for the duration of the drive to the best of your ability. You will also be using the automated lane keeping system in this drive.

- For participants in the **low automation condition (ALK only):**
    - During this drive you will be using automated lane keeping. Automated lane keeping systems keep the vehicle in the center of the lane so that you do not have to steer. Remember, putting your hands on the steering wheel and turning it even a small amount can turn off the automated lane keeping system. To avoid doing so inadvertently, please keep your hands off of the steering wheel unless you feel that you must take control of the vehicle to avoid an accident.
    - You will know that the automated lane keeping system is on if the green, nearly parallel lines are present in the dashboard. If this display is not present, the automated lane keeping system is not on.

- For participants in the **high automation condition (ALK and ACC):**
    - During this drive you will be using automated lane keeping and adaptive cruise control. Automated lane keeping systems keep the vehicle in the center of the lane so that you do not have to steer. Adaptive cruise control systems maintain a preset speed and adjust based on vehicles directly in front of you so that you do not have to accelerate or decelerate. Remember, putting your hands on the steering wheel and turning it even a small amount can turn off the automated lane keeping system. Similarly, if you press either the gas or brake pedal, the adaptive cruise control system will turn off. To avoid doing so inadvertently, please keep your hands off of the steering wheel and your foot off the pedals unless you feel that you must take control of the vehicle to avoid an accident.
    - You will know that the automated lane keeping system is on if the green, nearly parallel lines are present in the dashboard. Similarly, you will know

if the adaptive cruise control system is on if a car icon is present in the dashboard. If these displays are not present, the automated lane keeping system is not on.

During the drive you will again have the option to engage in a game on the side tablet screen. You are only to play the game if you feel it is safe to do so. Your primary task is to drive safely. A screen with a “Play” button will be presented and will remain on the screen unless you decide it is safe to play the game. If you press the “Play” button, the game will begin. Periodically, you will be presented questions on the tablet screen. You will know a question is being presented when the “Ready” button appears on the tablet. If you are ready to answer the question, you will simply press the “Ready” button and answer the question on the tablet. After you answer the question, the “Play” screen for Tetris will reappear. Do you have any questions?

Please complete the drive as safely as possible.

Completion

Upon completion, we will stop the driving scenario and present you with the final set of questionnaires to complete.

# APPENDIX E. SITUATION AWARENESS RATING TECHNIQUE (SART)

Based on the drive you just completed, please fill out the following questions:

Instability of Situation

| | 1 (1) | 2 (2) | 3 (3) | 4 (4) | 5 (5) | 6 (6) | 7 (7) |
|---|---|---|---|---|---|---|---|
| How changeable is the situation? Is the situation highly unstable and likely to change suddenly (High) or is it very stable and straightforward (Low)? (1) | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ |

Complexity of Situation

| | 1 (1) | 2 (2) | 3 (3) | 4 (4) | 5 (5) | 6 (6) | 7 (7) |
|---|---|---|---|---|---|---|---|
| How complicated is the situation? Is it complex with many interrelated components (High) or is it simple and straightforward (Low)? (1) | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ |

Variability of Situation

| | 1 (1) | 2 (2) | 3 (3) | 4 (4) | 5 (5) | 6 (6) | 7 (7) |
|---|---|---|---|---|---|---|---|
| How many variables are changing within the situation? Are there a large number of factors varying (High) or are there very few variables changing (Low)? (1) | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ |

Arousal

| | 1 (1) | 2 (2) | 3 (3) | 4 (4) | 5 (5) | 6 (6) | 7 (7) |
|---|---|---|---|---|---|---|---|
| How aroused are you in the situation? Are you alert and ready for activity (High) or do you have a low degree of alertness (Low)? (1) | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ |

Concentration of Attention

| | 1 (1) | 2 (2) | 3 (3) | 4 (4) | 5 (5) | 6 (6) | 7 (7) |
|---|---|---|---|---|---|---|---|
| How much are you concentrating on the situation? Are you concentrating on many aspects of the situation (High) or focused on only one (Low)? (1) | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ |

Division of Attention

| | 1 (1) | 2 (2) | 3 (3) | 4 (4) | 5 (5) | 6 (6) | 7 (7) |
|---|---|---|---|---|---|---|---|
| How much is your attention divided in the situation? Are you concentrating on many aspects of the situation (High) or focussed on only one (Low)? (1) | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ |

Spare Mental Capacity

| | 1 (1) | 2 (2) | 3 (3) | 4 (4) | 5 (5) | 6 (6) | 7 (7) |
|---|---|---|---|---|---|---|---|
| How much mental capacity do you have to spare in the situation? Do you have sufficient to attend to many variables (High) or nothing to spare at all (Low)? (1) | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ |

Information Quantity

| | 1 (1) | 2 (2) | 3 (3) | 4 (4) | 5 (5) | 6 (6) | 7 (7) |
|---|---|---|---|---|---|---|---|
| How much information have you gained about the situation? Have you received or understood a great deal of knowledge (High) or very little (Low)? (1) | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ |

Familiarity with Situation

| | 1 (1) | 2 (2) | 3 (3) | 4 (4) | 5 (5) | 6 (6) | 7 (7) |
|---|---|---|---|---|---|---|---|
| How familiar are you with the situation? Do you have a great deal of relevant experience (High) or is it a new situation (Low)? (1) | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ |

# APPENDIX F. TRUST IN AUTOMATION SCALE FOR INITIAL VALIDATION

Please rate the extent to which you agree or disagree with the following statements regarding the videos of the automated vehicle that you just watched.

| | 1 (1) | 2 (2) | 3 (3) | 4 (4) | 5 (5) | 6 (6) | 7 (7) |
|---|---|---|---|---|---|---|---|
| The automated vehicle is deceptive. (1) | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ |
| The automated vehicle behaves in an underhanded manner. (2) | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ |
| I am suspicious of the automated vehicle's intent, action, or outputs. (3) | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ |
| I am wary of the automated vehicle. (4) | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ |
| The automated vehicle's actions will have a harmful or injurious outcome. (5) | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ |
| I am confident in the automated vehicle. (6) | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ |
| The automated vehicle has integrity. (7) | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ |
| The automated vehicle is dependable. (8) | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| The automated vehicle is reliable. (9) | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ |
| I can trust the automated vehicle. (10) | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ |
| I am familiar with the automated vehicle. (11) | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ |
| The automated vehicle provides safety. (12) | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ |

# APPENDIX G. TRUST IN AUTOMATION SCALE FOR SIMULATOR VALIDATION

Based on the drive you just completed, please fill out the following questions where 1 = not at all and 7 = extremely.

| | 1 (1) | 2 (2) | 3 (3) | 4 (4) | 5 (5) | 6 (6) | 7 (7) |
|---|---|---|---|---|---|---|---|
| The system is deceptive. (1) | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ |
| The system behaves in an underhanded manner. (2) | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ |
| I am suspicious of the system's intent, action, or outputs. (3) | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ |
| I am wary of the system. (4) | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ |
| The system's actions will have a harmful or injurious outcome. (5) | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ |
| I am confident in the system. (6) | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ |
| The system provides security. (7) | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ |
| The system has integrity. (8) | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ |
| The system is dependable. (9) | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ |
| The system is reliable. (10) | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ |
| I can trust the system. (11) | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ |
| I am familiar with the system. (12) | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ | ❍ |

# APPENDIX H. NASA TLX DEFINITIONS

| Title | Endpoints | Description |
|---|---|---|
| Mental Demand | Low/High | How much mental and perceptual activity was required (e.g. thinking, deciding, calculating, remembering, looking, searcihing, etc.)? Was the task easy or demanding, simple or complex, exacting or forgiving? |
| Physical demand | Low/High | How much physical activity was required (e.g. pushing, pulling, turning, controlling, activating, ect.)? Was the task easy or demanding, slow or brisk, slack or strenuous, restful or laborious? |
| Temporal demand | Low/High | How much time pressure did you feel due to the rate or pace at which the tasks or task elements occurred? Was the pace slow and leisurely or rapid and frantic? |
| Performance | Good/Poor | How successful do you think you were accomplishing the goals of the task set by the experimenter (or yourself)? How satisfied were you with your performance in accomplishing these goals? |
| Effort | Low/High | How hard did you have to work (mentally and physically) to accomplish your level of performance? |
| Frustration level | Low/High | How insecure, discouraged, irritated, stressed and annoyed versus secure, gratified, content, relaxed and complacent did you feel during the task? |

# APPENDIX I. SIMULATOR VALIDATION DEBRIEF FORM

**Thanks and Introduction**
First of all, thank you for your participation in this experiment. We are members of Sonification Lab in the School of Psychology.

**Purpose of Experiment**
The purpose of this experiment was to investigate how situation awareness is affected by different levels of automation. In each of the two drives, we measured your eye movements, pupil size, driving performance, workload, awareness of the driving environment, trust in the automated system, and feelings toward the automated system.

**Meaning of Expected Results**
We expect that as automation increases participants' situation awareness, their knowledge of the driving environment, will worsen. We expect that analysis of eye movements, driving performance, workload, awareness of the driving environment, and trust will help identify what information participants attend to in highly automated vehicles and how it influences awareness of the driving environment. These results will be used to establish guidelines for the design of displays for automated driving.

**Confidentiality and Anonymity**
The results of your experiment will be used for only psychological study and never used for any other purposes. The data that is collected from you will be kept private to the extent required by law. To protect your privacy, your records will be kept under a code number rather than by name. Your records will be kept in locked files and only research staffs will be allowed to look at them. Your name and any other fact that might point to you will not appear when results of this study are presented or published. To make sure that this research is being carried out in the proper way, the Georgia Institute of Technology IRB will review study records. Again, your privacy will be protected to the extent required by law.

**Conclusion**
All of the experiment procedures are finished. We very much appreciate your efforts again.

**Contact Information**
For further information of this research, contact:

**Principal Investigator**
Dr. Bruce Walker (bruce.walker@psych.gatech.edu)

**Experimenters**

Zoe Becerra (zbecerra3@gatech.edu)
Sanghavi Gaddam (sanghavig@gatech.edu)
Sahar Ali (saharnazimali@gmail.com)
Brittany Noah (brittany.noah@gatech.edu)

# APPENDIX J. HOW TO USE THE SITUATIONAL TRUST SCALE FOR AUTOMATED DRIVING (STS-AD)

To use the STS-AD, present the items in the table below in the order that is presented here after participants experience a Level 3 or higher automated driving system. Items should be collected with a 7-point Likert scale ranging from (1 – completely disagree; 7 – completely agree).

After the data is collected, reverse score the Performance, Risk, and Judgement items (1 = 7; 2 = 6; 3 = 5; 5 = 3; 6 = 2; 7 = 1). Then, compute an average agreement score for the six items. This average score is then the total for the STS-AD.

| Item | Item Abbreviation |
|---|---|
| I trust the automation in this situation. | Trust |
| I would have performed better than the automation in this situation. (Reverse scored.) | Performance |
| In this situation, the automated vehicle performs good enough for me to engage in other activities (such as reading). | NDRT (non-driving related task) |
| The situation was risky. (Reverse scored.) | Risk |
| The automated vehicle made an unsafe judgement in this situation. (Reverse scored.) | Judgement |
| The automated vehicle reacted appropriately to the environment. | Reaction |

# APPENDIX K. HOW TO DEVELOP A SITUATIONAL TRUST SCALE

The following describes the process to develop a situational trust scale for a different context of automation use.

1. Identify key task characteristics through a task analysis. These key task characteristics will be used to develop the items for assessing situational trust.
2. Develop items that assess the external and internal variability of situational trust as described by Hoff and Bashir (2015).
3. Collect pilot data with a subset of your target sample using the situational trust scale and the Jian et al. 2000 scale for general trust in automation.
4. Evaluate the data using the following goals:
    a. Separability of the situational trust measure from the general trust measure.
        i. Compute a PCA with all of the Jian et al. 2000 items and the situational trust items in order to determine how many factors to extract for the EFA. Compute an EFA using eigenvalue limited extraction or limiting the number of factors based on the PCA results. Use the varimax rotation to help determine which items load onto which factors.
        ii. The situational trust items should load onto different factors than the general trust items. If this is not the case, the items may not be specific enough for the context. Revisit the items to ensure that

they specifically relate to the task characteristics experienced by users of the automation.

b. Factorial structure of the situational trust scale.

    i. Compute a PCA with only the situational trust items. This will determine how many factors to extract for the EFA. Compute an EFA using eigenvalue limited extraction or limiting the number of factors based on the PCA results.

    ii. The purpose of completing this analysis is to determine how many underlying factors there are in the situational trust scale. This will depend greatly on the number of items and the breadth of the items. It is possible to have multiple subscales measuring factors related to situational trust, especially if there are multiple items per factor identified by Hoff and Bashir.

c. Internal reliability

    i. Compute a Cronbach's alpha test to determine the internal reliability of the scale. Ideally, the alpha level should be 0.8 or higher. Consider removing items from the scale if removing them would improve the reliability of the scale.

    ii. The purpose of this evaluation is to determine if the items in the scale are related to each other – indicating that they are measuring the same construct. If from completing the factorial structure assessment, you determine that there are subscales (more than one factor extracted) then you can run this analysis once with all items and once with the items of each subscale.

## APPENDIX L. DESCRIPTIVE STATISTICS FOR THE STS-AD FROM THE SIMULATOR EVALUATION

| Automation Level | Item | Mean | Standard Deviation |
|---|---|---|---|
| **Low** | I trust the automation in this situation. | 5.49 | 1.392 |
| | I would have performed better than the automation in this situation. (Reverse scored.) | 5.09 | 1.345 |
| | In this situation, the automated vehicle performs good enough for me to engage in other activities (such as reading). | 5.42 | 1.454 |
| | The situation was risky. (Reverse scored.) | 3.96 | 1.758 |
| | The automated vehicle made an unsafe judgement in this situation. (Reverse scored.) | 6.07 | 1.009 |
| | The automated vehicle reacted appropriately to the environment. | 5.87 | 1.160 |
| **High** | I trust the automation in this situation. | 5.49 | 1.517 |
| | I would have performed better than the automation in this situation. (Reverse scored.) | 4.91 | 1.411 |
| | In this situation, the automated vehicle performs good enough for me to engage in other activities (such as reading). | 5.69 | 1.311 |

| | | | |
|---|---|---|---|
| | The situation was risky. (Reverse scored.) | 4.20 | 1.841 |
| | The automated vehicle made an unsafe judgement in this situation. (Reverse scored.) | 5.69 | 1.564 |
| | The automated vehicle reacted appropriately to the environment. | 5.84 | 1.107 |

# REFERENCES


Beller, J., Heesen, M., & Vollrath, M. (2013). Improving the Driver-Automation Interaction: An Approach Using Automation Uncertainty. *Human Factors*, *55*(6). https://doi.org/10.1177/0018720813482327

Chancey, E. T., Bliss, J. P., Yamani, Y., & Handley, H. A. H. (2017). Trust and the Compliance-Reliance Paradigm: The Effects of Risk, Error Bias, and Reliability on Trust and Dependence. *Human Factors*, *59*(3), 333–345. https://doi.org/10.1177/0018720816682648

Cotte, N., Meyer, J., & Coughlin, J. F. (2001). Older and Younger Drivers' Reliance on Collision Warning Systems. In *Proceedings of the Human Factors and Ergonomics Society 45th Annual Meeting* (pp. 277–280).

Cronbach, L. J. (1951). Coefficient Alpha and the Internal Structure of Tests. *Psychometrika*, *16*(3), 297–334.

Dixon, S. R., Wickens, C. D., & McCarley, J. S. (2007). On the independence of compliance and reliance: are automation false alarms worse than misses? *Human Factors*, *49*(4), 564–572. https://doi.org/10.1518/001872007X215656

Durso, F. T., & Dattel, A. R. (2004). SPAM: The real-time assessment of SA. In S. Banbury & S. Tremblay (Eds.), *A Cognitive approach to situation awareness: theory and application* (pp. 137–154). Aldershot, UK: Ashgate.

Ekman, F., Johansson, M., & Sochor, J. (2018). Creating appropriate trust in automated vehicle systems: A framework for HMI design. *IEEE Transactions on Human-Machine Systems*, *48*(1), 95–101. https://doi.org/10.1109/THMS.2017.2776209

Endsley, M. R., & Kaber, D. B. (1999). Level of automation effects on performance, situation awareness and workload in a dynamic control task. *Ergonomics*, *42*(3), 462–492. https://doi.org/10.1080/001401399185595

Fabrigar, L. R., & Wegener, D. T. (2012). *Exploratory Factor Analysis*. New York: Oxford University Press.

Frazier, M. L., Johnson, P. D., & Fainshmidt, S. (2013). Development and validation of a propensity to trust scale. *Journal of Trust Research*, *3*(2), 76–97. https://doi.org/10.1080/21515581.2013.820026

Gable, T. M., & Walker, B. N. (2013). *Georgia Tech Simulator Sickness Screening Protocol*. Atlanta.

Geels-Blair, K., Rice, S., & Schwark, J. (2013). Using System-Wide Trust Theory to Reveal the Contagion Effects of Automation False Alarms and Misses on Compliance and Reliance in a Simulated Aviation Task. *The International Journal of Aviation Psychology*, *23*(3), 245–266. https://doi.org/10.1080/10508414.2013.799355

Gold, C., Körber, M., Hohenberger, C., Lechner, D., & Bengler, K. (2015). Trust in automation – Before and after the experience of take-over scenarios in a highly automated vehicle. *Procedia Manufacturing*, *3*(3), 3025–3032. https://doi.org/10.1016/j.promfg.2015.07.847

Hancock, P. A., Jagacinski, R. J., Parasuraman, R., Wickens, C. D., Wilson, G. F., & Kaber, D. B. (2013). Human-automation interaction research: Past, present, and future. *Ergonomics in Design*, *21*(2), 9–14. https://doi.org/10.1177/1064804613477099

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index):

Results of Empirical and Theoretical Research. *Advances in Psychology*, *52*(C), 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9

Helldin, T., Falkman, G., Riveiro, M., & Davidsson, S. (2013). Presenting system uncertainty in automotive UIs for supporting trust calibration in autonomous driving. *Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications - AutomotiveUI '13*, *5*(October), 210–217. https://doi.org/10.1145/2516540.2516554

Hergeth, S., Lorenz, L., Vilimek, R., & Krems, J. F. (2016). Keep Your Scanners Peeled: Gaze Behavior as a Measure of Automation Trust During Highly Automated Driving. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *58*(3), 509–519. https://doi.org/10.1177/0018720815625744

Hoff, K. A., & Bashir, M. (2015). Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust . *Human Factors: The Journal of the Human Factors and Ergonomics Society* , *57*(3). https://doi.org/10.1177/0018720814547570

Jian, J.-Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an Empirically Determined Scale of Trust in Automated System. *International Journal of Cognitive Ergonomics*, *4*(1), 53. https://doi.org/10.1207/S15327566IJCE0401_04

Körber, M. (2019). Theoretical considerations and development of a questionnaire to measure trust in automation. *Advances in Intelligent Systems and Computing*, *823*(March), 13–30. https://doi.org/10.1007/978-3-319-96074-6_2

Körber, M., Baseler, E., & Bengler, K. (2018). Introduction matters: Manipulating trust in automation and reliance in automated driving. *Applied Ergonomics*, *66*, 18–31. https://doi.org/10.1016/j.apergo.2017.07.006

Large, D. R., Burnett, G., Morris, A., Muthumani, A., & Matthias, R. (2018). A longitudinal simulator study to explore drivers' behaviour during highly-automated driving. In *Advances in Intelligent Systems and Computing*. https://doi.org/10.1007/978-3-319-60441-1_57

Lee, J. D., & See, K. A. (2004). Trust in Automation: Designing for Appropriate Reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50.30392

Li, M., Holthausen, B. E., Stuck, R. E., & Walker, B. N. (2019). No risk no trust: Investigating perceived risk in highly automated driving. *Proceedings - 11th International ACM Conference on Automotive User Interfaces and Interactive Vehicular Applications, AutomotiveUI 2019*, 177–185. https://doi.org/10.1145/3342197.3344525

Marsh, S., & Dibben, M. R. (2005). The role of trust in information science and technology. *Annual Review of Information Science and Technology*, *37*(1), 465–498. https://doi.org/10.1002/aris.1440370111

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An Integrative Model of Organizational Trust. *The Academy of Management Review*, *20*(3), 709–734. Retrieved from http://www.jstor.org/stable/258792

Merritt, S. M., & Ilgen, D. R. (2008). Not All Trust Is Created Equal: Dispositional and History-Based Trust in Human-Automation Interactions. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *50*(2), 194–210. https://doi.org/10.1518/001872008X288574

National Highway Transportation Administration. (n.d.). Automated Vehicles for Safety.

Retrieved March 5, 2019, from https://www.nhtsa.gov/technology-innovation/automated-vehicles-safety

Noah, B. E. (2018). *Understanding Automation Handoff Impacts on Workload and Trust when Mitigated by Reliability Displays*. Georgia Institute of Technology.

Noah, B. E., Gable, T. M., Chen, S.-Y., Singh, S., & Walker, B. N. (2017). Development and Preliminary Evaluation of Reliability Displays for Automated Lane Keeping. *Proceedings of the International Conference on Automotive User Interfaces and Interactive Vehicular Applications - AutomotiveUI '17*, 202–208. https://doi.org/10.1145/3122986.3123007

Noah, B. E., Gable, T. M., & Walker, B. N. (2016). Ordinal magnitude scaling for automated lane keeping displays. *AutomotiveUI 2016 - 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, Adjunct Proceedings*. https://doi.org/10.1145/3004323.3004341

Noah, B. E., & Walker, B. N. (2017). Trust calibration through reliability displays in automated vehicles. In *HRI'17 Proceedings of the Companion of the ACM/IEEE International Conference on Human-Robot Interaction* (pp. 361–362). https://doi.org/10.1145/3029798.3034802

Nunnally, J. C. (1978). *Psychometric theory*. New York: McGraw Hill.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use , misuse , disuse , abuse. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *39*(2), 230–253. https://doi.org/10.1518/001872097778543886

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and*

*Cybernetics Part A:Systems and Humans.*, *30*(3), 286–297. https://doi.org/10.1109/3468.844354

Pituch, K. A., & Stevens, J. P. (2016). *Applied Multivariate Statistics for the Social Sciences* (6th ed.). New York: Routledge.

Rice, S, & Keller, D. (2009). Automation reliance under time pressure. *Cognitive Technology*, (1993), 36–44. https://doi.org/10.1518/001872007X215656

Rice, Stephen. (2009). Examining Single- and Multiple-Process Theories of Trust in Automation. *Journal of General Psychology*, *136*(3), 303–319. https://doi.org/10.3200/GENP.136.3.303-322

Rice, Stephen, & Geels, K. (2010). Using System-Wide Trust Theory to Make Predictions About Dependence on Four Diagnostic Aids. *The Journal of General Psychology*, *137*(4), 362–375. https://doi.org/10.1080/00221309.2010.499397

Rice, Stephen, & McCarley, J. S. (2011). Effects of response bias and judgment framing on operator use of an automated aid in a target detection task. *Journal of Experimental Psychology: Applied*, *17*(4), 320–331. https://doi.org/10.1037/a0024243

SAE International. (2014). Taxonomy and Definitions for Terms Related to On-Road Motor Vehicle Automated Driving Systems. *SAE International*, *J3016*, 1–12. https://doi.org/10.4271/J3016_201609

Sawyer, B. D., Seppalt, B., Mehler, B., & Reimer, B. (2017). Trust Impacts Driver Glance Strategy in Multitasking (pp. 1441–1442). https://doi.org/10.1177/1541931213601845

Seppelt, B. D., & Lee, J. D. (2007). Making adaptive cruise control (ACC) limits visible.

*International Journal of Human Computer Studies*, *65*(3), 192–205. https://doi.org/10.1016/j.ijhcs.2006.10.001

Stuck, R. E., & Rogers, W. A. (2018). Older Adults' Perceptions of Supporting Factors of Trust in a Robot Care Provider. *Journal of Robotics*, *2018*. https://doi.org/10.1155/2018/6519713

Stuck, R. E., Tomlinson, B. J., & Walker, B. N. (n.d.). The Importance of Incorporating Risk into the Study of Human-Automation Trust. *Manuscript in Publication*.

Taylor, R. M. (1990). Situational awareness rating technique (SART): The development of a tool for aircrew systems design. In *AGARD, Situational Awareness in Aerospace Operations* (pp. 90-28972 23-53).

The HART Group. (2011). *Automation in the Cockpit: Toward a Human-Automation Relationship Taxonomy* (Vol. Interim Re).

US Department of Transportation. (2018). *Preparing for the Future of Transportation*. https://doi.org/DOT-OST-2018-0149

Verberne, F. M. F., Ham, J., & Midden, C. J. H. (2012). Trust in smart systems: Sharing driving goals and giving information to increase trustworthiness and acceptability of smart systems in cars. *Human Factors*, *54*(5), 799–810. https://doi.org/10.1177/0018720812443825

Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology*, *52*, 113–117. https://doi.org/10.1016/j.jesp.2014.01.005

Wickens, C. (2018). Automation Stages & Levels, 20 Years After. *Journal of Cognitive*

*Engineering and Decision Making*, *12*(1), 35–41. https://doi.org/10.1177/1555343417727438

Wickens, C. D., Li, H., Santamaria, A., Sebok, A., & Sarter, N. B. (2010). Stages and Levels of Automation: An Integrated Meta-analysis. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *54*(4), 389–393. https://doi.org/10.1177/154193121005400425

Wickens, C. D., Tsang, P. S., & Benel, R. A. (1979). The Allocation of Attentional Resources in a Dynamic Environment. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. https://doi.org/https://doi.org/10.1177/1071181379023001131

Wojton, H., Lane, S., & Porter, D. (2019). Initial Validation of the Trust of Automated Systems Test (TOAST). In *Presentation at the Human Factors Engineering Technical Advisory Group*. Aberdeen Proving Ground, MD.

Yamada, K., & Kuchar, J. K. (2006). Preliminary study of behavioral and safety effects of driver dependence on a warning system in a driving simulator. *IEEE Transactions on Systems, Man, and Cybernetics Part A:Systems and Humans*, *36*(3), 602–610. https://doi.org/10.1109/TSMCA.2006.871646